Measuring the dynamics of information processing on a local scale in time and space

Joseph T. Lizier

Abstract Studies of how information is processed in natural systems, in particular in nervous systems, are rapidly gaining attention. Less known however is that the local dynamics of such information processing in space and time can be measured. In this chapter, we review the mathematics of how to measure local entropy and mutual information values at specific observations of time-series processes. We then review how these techniques are used to construct measures of local information storage and transfer within a distributed system, and we describe how these measures can reveal much more intricate details about the dynamics of complex systems than their more well-known "average" measures do. This is done by examining their application to cellular automata, a classic complex system, where these local information profiles have provided quantitative evidence for long-held conjectures regarding the information transfer and processing role of gliders and glider collisions. Finally, we describe the outlook in anticipating the broad application of these local measures of information processing in computational neuroscience.

1 Introduction

Analysis of directed information transfer between variables in time-series brain imaging data and models is currently gaining much attention in neuroscience. Measures of information transfer have been computed, for example, in fMRI measurements in the human visual cortex between average signals at the regional level [38] and between individual voxels [8], as well as between brain areas of macaques from local field potential (LFP) time-series [48]. A particularly popular topic in this domain is the use of information transfer measures to infer effective network connectivity between variables in brain-imaging data [39, 91, 49, 88, 69, 54, 63], as well as studying modulation of connection strength with respect to an underlying task

Joseph T. Lizier
CSIRO Computational Informatics, Marsfield, Australia, e-mail: joseph.lizier@csiro.au

[94]. Furthermore, measures of information transfer are used to reveal differences between healthy and diseased states in neural data (e.g. for EEG measurements of epilipsy patients in [10]) and in models (e.g. for Parkinson's disease in [43]).

Much of this work quantifies information transfer from a source variable to a target variable using the information-theoretic measure known as the *transfer entropy* [82], or its equivalent under linear-Gaussian conditions, the Granger causality [28]. This information-theoretic approach to studying directed interactions in neural systems can be viewed as part of a more broad effort to study distributed computation in complex systems in terms of how information is stored, transfered and modified (e.g. [59, 60, 62]). The approach is highly appropriate in computational neuroscience, and indeed for complex systems in general, because:

- these concepts of computation are meaningful and well-understood (e.g. information transfer as reflecting directed coupling between two variables, information storage as predictability or structure in a time-series process);
- the quantities measured (e.g. transfer entropy for measuring information transfer) are well-defined and can be measured on any type of time-series data (continuous or discrete-valued);
- the quantities are at heart model-free (in contrast to the Granger causality linearisation)[1] and detect non-linear interactions and structure; and
- distributed computation is the language in which dynamics are often described in neuroscience (e.g. "the brain represents and processes information in a distributed fashion and in a dynamical way" [27]) and complex systems in general (e.g. claims that small-world structures have "maximum capability to store, process and transfer information" [42]).

Now, such work on distributed computation to date typically focuses on the (time) *average* information transfer, which is how the transfer entropy and other information-theoretic measures are traditionally defined. Yet the *dynamics* of transfer from a source to a target can also be quantified at individual observations or configurations of the variables using the *local* transfer entropy [59]. Such local measures can be defined for any traditional information-theoretic variable, including for related measures of information storage and processing (e.g. [62]). To be explicit, local information-theoretic measures characterise the information attributed with specific measurements x and y of variables X and Y, rather than the average information associated with these variables.

This local perspective can reveal dynamical structure that the average cannot. Applied to time-series data, local measures tell us about the *dynamics* of information in the system, since they vary with the specific observations in time, and local values are known to reveal more details about the system than the averages alone [16, 83, 84]. To be specific, a measured average of transfer entropy does not tell us about how the directed relationship between two variables fluctuates through time, how different specific source states may be more predictive of a target than other states, or how coupling strength may relate to changing underlying experimental conditions.

[1] This also contrasts with dynamic causal modeling, a model-based approach that compares a set of a priori defined neural models and tests how well they explain the experimental data [25].

Indeed, the ability to investigate *time-series dynamics* of distributed computation in complex systems provides an important connection from information theory to *dynamical systems theory* or *non-linear time-series analysis* (e.g. see [81, 41]). We use the term **information dynamics** to describe the study of distributed computation in complex systems in terms of how information is stored, transferred and modified [59, 60, 62]. The word *dynamics* is a key component of this term, referring to both:

1. That we study the dynamic *state updates* of variables in the system, decomposing information in the measurement of a variable in terms of information from that variable's own past (information storage), information from other variables (information transfer) and how those information sources are combined (information modification);
2. That we study *local information-theoretic measures* for each of these variables, quantifying the dynamics of these operations in time and space.

In this chapter, we review how such local information-theoretic measurements can be made, and describe how they are used to define local measures of information storage and transfer in distributed computation in complex systems. We begin by describing the relevant information-theoretic concepts in Sect. 2, before providing a detailed presentation of how local information-theoretic measures are defined in Sect. 3. We then provide an overview of our framework for information dynamics in Sect. 4, describing the measures used for information storage and transfer, and how they can be localised within a system in space and time using the techniques of Sect. 3. Next, we review in Sect. 5 the application of these local measures of computation to cellular automata, a simple discrete dynamical model which is known to exhibit complex behaviour and emergent coherent structures (known as particles or gliders) resembling coherent waves in neural dynamics [27]. This application demonstrates the utility of these local measures of information storage and transfer, by providing key insights into the dynamics of cellular automata, including demonstrating evidence for long-held conjectures regarding the computational role of the emergent structures (e.g. gliders as information transfer entities). Most importantly, the local measures are shown to provide insights into the dynamics of information in the system that are simply not possible to obtain with traditional averaged information-theoretic methods.

We finish the chapter by describing in Sect. 6 further such insights into the dynamics of information that have since been obtained with these local measures for other systems. For example, the measures have revealed coherent information cascades spreading across flocks (or swarms) [92] and in modular robots [57], in analogy to the aforementioned gliders in cellular automata. They have also demonstrated the key role of information transfer in network synchronization processes, in particular in indicating when a synchronized state has been "computed" but not yet obviously reached [9]. Just like the cellular automata examples, these demonstrate the ability of local information dynamics to reveal how the computation in a system unfolds in time, and the dynamics of how separate agents or entities interact to achieve a collective task. Crucially, they allow one to answer meaningful questions about the information processing in a system, in particular: "*when* and *where* is informa-

tion transferred in the brain during cognitive tasks?", and we describe a preliminary study where this precise question is explored using fMRI recordings during a button pressing task. As such, we demonstrate that local information dynamics enables whole new lines of inquiry which were not previously possible in computational neuroscience or other fields.

2 Information-theoretic preliminaries

To quantify the information dynamics of distributed computation, we first look to information theory (e.g. see [85, 13, 65]) which has proven to be a useful framework for the design and analysis of complex self-organized systems, e.g. [14, 77, 78, 66]. In this section, we give a brief overview of the fundamental quantities which will be built on in exploring local information dynamics in the following sections.

The fundamental quantity of information theory is the **Shannon entropy**, which represents the average uncertainty associated with any measurement x of a random variable X (logarithms are taken by convention in base 2, giving units in bits):

$$H(X) = -\sum_{x} p(x) \log_2 p(x). \tag{1}$$

The uncertainty $H(X)$ associated with such a measurement is equal to the information required to predict it (see self-information below).

The Shannon entropy was originally derived following an axiomatic approach. This is important because it gives primacy to desired properties over candidate measures, rather than retrospectively highlighting properties of an appealing candidate measure. It shifts the focus of any arguments over the form of measures onto the more formal ground of selecting which axioms should be satisfied. This is particularly useful where a set of accepted axioms can uniquely specify a measure (as in the cases discussed here). We highlight the axiomatic approach here because it has persisted in later developments in information theory, in particular for the local measures we discuss in Sect. 3 (as well as more recently in debate over measures of information redundancy [95, 35, 53]).

So, the Shannon entropy was derived as the unique formulation (up to the base of the logarithm) satisfying a certain set of properties or axioms [85] (with property labels following [76]):

- **continuity** with respect to the underlying probability distribution function $p(x)$ (PDF). This sensibly ensures that small changes in $p(x)$ only lead to small changes in $H(X)$.
- **monotony**: being a monotonically increasing function of the number of choices n for x when each choice x_i is equally likely (with probability $p(x_i) = 1/n$). In Shannon's words, this desirable because: "With equally likely events there is more choice, or uncertainty, when there are more possible events" [85].

- **grouping**: "If a choice (can) be broken down into two successive choices, the original H should be the weighted sum of the individual values of H" [85]. That is to say, "H is independent of how the process is divided into parts" [76]. This is crucial because the intrinsic uncertainy we measure for the process should not depend on any subjectivity in how we divide up the stages of the process to be examined.

Further, note that the Shannon entropy for a measurement can be interpreted as the minimal average number of bits required to encode or describe its value without losing information [65, 13].

The **joint entropy** of two random variables X and Y is a generalization to quantify the uncertainty of their joint distribution:

$$H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y). \tag{2}$$

The **conditional entropy** of X given Y is the average uncertainty that remains about x when y is known:

$$H(X \mid Y) = -\sum_{x,y} p(x,y) \log_2 p(x \mid y). \tag{3}$$

The conditional entropy for a measurement of X can be interpreted as the minimal average number of bits required to encode or describe its value without losing information, given that the receiver of the encoding already knows the value of Y. The previous quantities are related by the following chain rule:

$$H(X,Y) = H(X) + H(Y \mid X). \tag{4}$$

The **mutual information** (MI) between X and Y measures the average reduction in uncertainty about x that results from learning the value of y, or vice versa:

$$I(X;Y) = -\sum_{x,y} p(x,y) \log_2 \frac{p(x \mid y)}{p(x)} \tag{5}$$

$$= H(X) - H(X \mid Y). \tag{6}$$

The MI is symmetric in the variables X and Y. The mutual information for measurements of X and Y can be interpreted as the average number of bits *saved* in encoding or describing X given that the receiver of the encoding already knows the value of Y, in comparison to the encoding of X without the knowledge of Y. These descriptions of X with and without the value of Y are both minimal without losing information. Note that one can compute the *self-information* $I(X;X)$, which is the average information required to predict the value of X, and is equal to the uncertainty $H(X)$

associated with such a measurement.

The **conditional mutual information** between X and Y given Z is the mutual information between X and Y when Z is known:

$$I(X;Y \mid Z) = -\sum_{x,y,z} p(x,y,z) \log_2 \frac{p(x \mid y,z)}{p(x \mid z)} \tag{7}$$

$$= H(X \mid Z) - H(X \mid Y,Z). \tag{8}$$

One can consider the MI from two variables Y_1, Y_2 jointly to another variable X, $I(X;Y_1,Y_2)$, and using (4), (6) and (8) decompose this into the information carried by the first variable plus that carried by the second conditioned on the first:

$$I(X;Y_1,Y_2) = I(X;Y_1) + I(X;Y_2 \mid Y_1). \tag{9}$$

Of course, this *chain rule* generalises to multivariate $\mathbf{Y}$ of dimension greater than two.

Note that a conditional MI $I(X;Y \mid Z)$ may be either larger or smaller than the related unconditioned MI $I(X;Y)$ [65]. The conditioning removes information *redundantly* held by the source Y and the conditioned variable Z about X (e.g. if both Y and Z were copies of X). Furthermore, the conditioning also includes *synergistic* information about X which can only be decoded with knowledge of both the source Y and conditioned variable Z (e.g. where X is the result of an exclusive-OR or XOR operation from Y and Z). These components *cannot* be teased apart with traditional information-theoretic analysis; the partial information decomposition approach was introduced for this purpose [95] (and see also [35, 32, 53]).

We now move on to consider measures of information in time-series processes X of the random variables $\{\ldots X_{n-1}, X_n, X_{n+1} \ldots\}$ with process realisations $\{\ldots x_{n-1}, x_n, x_{n+1} \ldots\}$ for countable time indices n. We refer to measures which consider how the information in variable X_n is related to previous variables, e.g. X_{n-1}, of the process or other processes as measures of **information dynamics**.

The **entropy rate** is defined by [13]:

$$H'_\mu(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{10}$$

$$= \lim_{n\to\infty} \frac{1}{n} H(\mathbf{X}_n^{(n)}), \tag{11}$$

(where the limit exists) where we have used $\mathbf{X}_n^{(k)} = \{X_{n-k+1}, \ldots, X_{n-1}, X_n\}$ to denote the k consecutive variables of X up to and including time step n. This quantity describes the limiting rate at which the entropy of n consecutive measurements of X grow with n. A related definition is given by:[2]

[2] Note that we have reversed the use of the primes in the notation from [13], in line with [14].

$$H_\mu(X) = \lim_{n\to\infty} H\left[X_n \mid X_1, X_2, \ldots, X_{n-1}\right] \tag{12}$$

$$= \lim_{n\to\infty} H\left[X_n \mid \mathbf{X}_{n-1}^{(n-1)}\right]. \tag{13}$$

Cover and Thomas [13] point out that these two quantities correspond to two subtly different notions: the first is something of an average per symbol entropy, while the second is a conditional entropy of the last random variable given the past. These authors go on to demonstrate that for stationary processes X, the limits for the two quantities $H'_\mu(X)$ and $H_\mu(X)$ exist (i.e. the average entropy rate converges) and are equal.

For our purposes in considering information dynamics, we are interested in the latter formulation $H_\mu(X)$, since it explicitly describes how one random variable X_n is related to the previous instances $\mathbf{X}_{n-1}^{(n-1)}$. For practical usage, we are particularly interested in estimation of $H_\mu(X)$ with finite-lengths k, and in estimating it regarding the information at different time indices n. That is to say, we use the notation $H_\mu(X_{n+1}, k)$ to describe the conditional entropy in X_{n+1} given $\mathbf{X}_n^{(k)}$:

$$H_\mu(X_{n+1}, k) = H\left[X_{n+1} \mid \mathbf{X}_n^{(k)}\right]. \tag{14}$$

Of course, letting $k = n$ and joining (13) and (14) we have $\lim_{n\to\infty} H_\mu(X_{n+1}, n) = H_\mu(X)$.

3 Local information theoretic measures

In this section, we describe how one may obtain *local* information measures with reference to their more well-known *average* information-theoretic counterparts. Local information-theoretic measures characterise the information attributed with specific measurements x and y of variables X and Y, rather than the average information associated with these variables. Local values within a global average are known to provide important insights into the dynamics of nonlinear systems [16].

We begin by defining local values of the entropy and conditional entropy (Shannon information content values) in Sect. 3.1, and then describe local mutual information and conditional mutual information in Sect. 3.2. Next, in Sect. 3.3 we consider the meaning and properties of these local values where where X and Y are time-series processes and local information-theoretic measures characterise the information attributed at each local point in time in these series. Finally, we describe in Sect. 3.4 the mechanics of how these local information-theoretic measures can be practically quantified, using various types of estimators.

Before beginning, we note that such local information-theoretic measures have been used (with less explicit presentation) in various earlier studies in complex systems science, e.g. for the local excess entropy [83], the local statistical complexity [83, 84], and the local information [36]. Yet relatively little exploration has been

made into the dynamics of these local information measures in complex systems, and certainly none had been made into the local dynamics of information storage, transfer and modification, as we will review in Sect. 4.

3.1 Shannon information content and its meaning

The **Shannon information content** or **local entropy** of an outcome x of measurement of the variable X is [65]:

$$h(x) = -\log_2 p(x). \tag{15}$$

Note that by convention we use lower-case symbols to denote local information-theoretic measures throughout this chapter. The Shannon information content can be shown to be the unique formulation (up to the base of the logarithm) satisfying the following properties [1]:

- **grouping**: $h(p_1(x_1) \times p_2(x_2)) = h(p_1(x_1)) + h(p_2(x_2))$, where $h(p(x)) = -\log_2 p(x) = h(x)$, and p_1 and p_2 (both satisfying $0 < p \leq 1$) can be interpreted as representing the probabilities of two independent events x_1 and x_2;
- **monotonically decreasing** with $p(x)$; and
- **continuity** with $p(x)$.

Note that these three properties map directly to the three properties for the (average) Shannon entropy (see Sect. 2). Also, noting that this quantity is also equivalent to a *local self-information*, it can also be derived (see [22, Chapter 2]) by starting with the local mutual information (see Sect. 3.2).

Now, the quantity $h(x)$ is simply the information content attributed to the specific symbol x, or the information required to predict or uniquely specify that specific value. Less probable outcomes x have higher information content than more probable outcomes, and we have $h(x) \geq 0$. Specifically, the Shannon information content of a given symbol x is the *code-length* for that symbol in an optimal encoding scheme for the measurements X, i.e. one that produces the minimal expected code length.[3]

In this light, one views the Shannon entropy as the "entropy of an ensemble" [65] of the outcomes x of the random variable X, with probabilities p defined over the alphabet A_x of possible outcomes. That is, $H(X)$ is the *average* or *expectation value* of the Shannon information content for each symbol $x \in A_x$ (compare to (1)):

$$H(X) = \sum_x p(x)h(x), \tag{16}$$

$$= \langle h(x) \rangle . \tag{17}$$

[3] Note that this "optimal code-length" may specify non-integer choices; full discussion of the implications of this, practical issues in selecting integer code-lengths, and block-coding optimisations are contained in [13, Chapter 5].

As we will see, each average information-theoretic measure is an average over its associated local quantity.

In the mathematics above, we see the average or expectation value as being taken over each symbol $x = m$ (where $m \in \{0, \ldots, M-1\}$ without loss of generality for some M discrete symbols). We can also view it however as being an average over each observation or measurement x_i (where i is a measurement index) of X that we used to construct our probability distribution function $p(x)$. To do this, we start from the operational definition of the PDF for each symbol: $p(x = m) = \frac{c(x=m)}{N}$, where $c(x = m)$ is the count of observations of the symbol m out of the N total observations. To precisely compute this probability, the ratio should be composed over all realisations of the observed variables (as described in [83]); realistically however, estimates will be made from a finite number of observations N. We then re-write (1) using this definition:

$$H(X) = -\sum_m \frac{c(x=m)}{N} \log_2 p(x=m), \tag{18}$$

and then further expand using the identity $c(x = m) = \sum_{g=1}^{c(x=m)} 1$:

$$H(X) = -\sum_m \sum_{g=1}^{c(x=m)} \frac{1}{N} \log_2 p(x=m). \tag{19}$$

This leaves a double sum running over i. each actual observation g, ii. for each possible observation $x = m$. This is equivalent to a single sum over all N observations x_i, $i = 1 \ldots N$, giving:

$$H(X) = -\frac{1}{N} \sum_{i=1}^{N} \log_2 p(x_i), \tag{20}$$

$$= \langle h(x_i) \rangle_i, \tag{21}$$

as required. To reiterate, we refer to $h(x_i)$ as a *local* entropy because it is defined locally for each observation x_i.

At this point, we note that the above derivation shows that the PDF $p(x)$ for the local value $h(x)$ is evaluated at a *specific* local observation x, but the function p is defined using *all* of the relevant observations. This is a subtle point - the *evaluation* of p is *local* to the observation x, but we need other observations to define the function p in order to make this evaluation. We revisit this concept when we consider time-series processes in Sect. 3.3.

Now, we note that one can also define **conditional Shannon information content** (or **local conditional entropy**) [65]:

$$h(x \mid y) = -\log_2 p(x \mid y), \tag{22}$$

and that these quantities satisfy the chain rule in alignment with their averages:

$$h(x,y) = h(y) + h(x \mid y). \tag{23}$$

In this way, we see that the information content of a joint quantity (x,y) is the code length of y plus the code length of x given y. Finally, we note that this quantity is also referred to as *conditional self-information* and can also be derived (see [22, Chapter 2]) by starting with the local conditional mutual information (see Sect. 3.2).

3.2 Local mutual information and conditional mutual information

Next, we consider localisations of the mutual information. One way to think about this quantity is to build the **local mutual information** directly from Shannon information content or local entropy measures, in alignment with its average definition, i.e.:

$$i(x;y) = h(x) - h(x \mid y), \tag{24}$$

$$= \log_2 \frac{p(x \mid y)}{p(x)}. \tag{25}$$

In this way, we see that the local mutual information is the difference in code lengths between coding the value x in isolation (under the optimal encoding scheme for X), or coding the value x given y (under the optimal encoding scheme for X given Y). In other words, this quantity captures the coding "cost" for x in not being aware of the value y. Similarly, the **local conditional mutual information** can be constructed as:

$$i(x;y \mid z) = h(x \mid z) - h(x \mid y,z), \tag{26}$$

$$= \log_2 \frac{p(x \mid y,z)}{p(x \mid z)}. \tag{27}$$

Here, we see that the local conditional mutual information is the difference in code lengths (or coding cost) between coding the value x given z (under the optimal encoding scheme for X given Z), or coding the value x given both y and z (under the optimal encoding scheme for X given Y and Z).

More formally however, Fano [22, ch. 2] set out to quantify "the amount of information provided by the occurrence of the event represented by y_i about the occurrence of the event represented by x_i." He derived the local mutual information $i(x;y)$ (25) to capture this concept, as well as the local conditional mutual information $i(x;y \mid z)$ (27), directly from the following four postulates:

- **once-differentiability** with respect to the underlying probability distribution functions $p(x)$ and $p(x \mid y)$;
- **identical mathematical form for the conditional MI and local conditional MI**, only with $p(x)$ replaced by $p(x \mid z)$ and $p(x \mid y)$ replaced by $p(x \mid y,z)$;
- **additivity** for the information provided by y and z about x, i.e.: $i(\{y,z\};x) = i(y;x) + i(z;x \mid y)$;

- **separation for independent ensembles** XY and UV, i.e. where we have $p(x,y,u,v) = p(x,y)p(u,v)$ then we must have $i(\{x,u\};\{y,v\}) = i(x;y) + i(u;v)$.

Crucially, Fano's derivation means that $i(x;y)$ and $i(x;y \mid z)$ are *uniquely specified*, up to the base of the logarithm.

Of course, we have $I(X;Y) = \langle i(x;y) \rangle$ and $I(X;Y \mid Z) = \langle i(x;y \mid z) \rangle$ as per the averaged entropy quantities in the previous section. It is particularly interesting that Fano made the derivation for local mutual information directly, and only computed the averaged quantity as a result of that. This contrasts with contemporary perspectives which generally give primary consideration to the averaged quantity. (This is not the case however in natural language processing for example, where the local MI is commonly used and known as the point-wise mutual information, e.g. [68]).

We also note that $i(x;y)$ is symmetric in x and y (like $I(X;Y)$), though this was not explicitly built into the above postulates.

Next, consider that the local MI and conditional MI values may be either positive or negative, in contrast to the local entropy which cannot take negative values. Positive values are fairly intuitive to understand: the local mutual information in (25) is positive where $p(x \mid y) > p(x)$, i.e. knowing the value of y increased our expectation of (or positively informed us about) the value of the measurement x. The existence of negative values is often a concern for readers unfamiliar with the concept, however they too are simple to understand. Negative values simply occur in (25) where $p(x \mid y) < p(x)$, i.e. knowing about the value of y actually changed our belief $p(x)$ about the probability of occurrence of the outcome x to a smaller value $p(x \mid y)$, and hence we considered it less likely that x would occur when knowing y than when not knowing y, in a case were x nevertheless occurred. As an example, consider the probability that it will rain today, $p(\texttt{rain} = 1)$, and the probability that it will rain given that the weather forecast said it would not, $p(\texttt{rain} = 1 \mid \texttt{rain\_forecast} = 0)$. Being generous to weather forecasters for a moment, let's say that $p(\texttt{rain} = 1 \mid \texttt{rain\_forecast} = 0) < p(\texttt{rain} = 1)$, so we would have $i(\texttt{rain} = 1; \texttt{rain\_forecast} = 0) < 0$, because we considered it less likely that rain would occur today when hearing the forecast than without the forecast, in a case where rain nevertheless occurred. These negative values of MI are actually quite meaningful, and can be interpreted as there being negative information in the value of y about x. We could also interpret the value y as being *misleading* or *misinformative* about the value of x, because it had *lowered* our expectation of observing x prior to that observation being made in this instance. In the above example, the weather forecast was misinformative about the rain today. One can also view the negative values using (24), seeing that $i(x;y)$ is negative where knowing y increased the uncertainty about x.

Importantly, these local measures always average to give a non-negative value. Elaborating on an example from Cover and Thomas [13, p.28], "in a court case, specific new evidence" y "might increase uncertainty" about the outcome x, "but on the average evidence decreases uncertainty". Similarly, in our above example, while the weather forecast might misinform us about the rain on a particular day, on average the weather forecast will provide positive (or at least zero!) information.

Finally, we note that the local mutual information $i(x;y)$ measures we consider here are distinct from partial localization expressions, i.e. the partial mutual information or specific information $I(x;Y)$ [18], which consider information contained in specific values x of one variable X about the other (unknown) variable Y. Crucially, there are two valid approaches to measuring partial mutual information, one which preserves the additivity property and one which retains non-negativity [18]. As described above however, there is only one valid approach for the fully local mutual information $i(x;y)$ (and see further discussion in [56]).

3.3 Local information measures for time series

Now, consider X_n, Y_n and Z_n as the variables of time-series processes X, Y and Z with specific measurements (x_n, y_n, z_n) at each time point $n = 1, \ldots, N$ (though the specific time interval is arbitrary).

The local information-theoretic measures, e.g. $i(x_n;y_n)$, then characterise the information attributed *at each local point in time* in these series. Furthermore, where $\mathbf{X}$ is a multivariate spatiotemporal series with measurements $x_{i,n}$ at spatial points i for each time n, then local information-theoretic measures, e.g. $i(x_{i,n};x_{i,n+1})$, characterise the information attributed at each local spatiotemporal point in the series, and one can form spatiotemporal *profiles* of the information characteristics. Such local characterisation is what we mean by the local measures being useful for studying the *dynamics* of information in space and time. We shall explore examples of such dynamics in the next sections.

As described earlier for $h(x)$, computing a local measure requires evaluating the probability distribution function (PDF) $p(x)$ for the given local observation x, however the PDF itself must be defined using all of the relevant observations of the variable X. Furthermore, where X is a time series, it is clear that the observations to construct the PDF for evaluating $p(x_n)$ at x_n are not local in time to that observation x_n. We must carefully consider which parts of the time series X are used to construct the PDF – one should select observations across which the time series is *stationary* or in the same phase of a *cyclostationary* process when constructing PDFs for information-theoretic functions.

Often, this may mean using a sliding window technique to construct the PDF – i.e. to evaluate $p(x_n)$ we may use observations $\{x_{n-T}, \ldots, x_{n+T}\}$ (for some T) to construct the PDF, assuming that the time series is stationary over that time-interval. While one would wish to maximise the size of the time-window in order to have many samples to estimate the PDF, this must be balanced against these stationarity considerations.

An alternate *ensemble approach* may be to sample many repeat time series X_i (where i is an instance, trial or realisation index of the time-series) with measurements $x_{i,n}$, where stationarity is assumed at fixed time points n over all samples i. In this case, $p(x_{i,n})$ is constructed for each $x_{i,n}$ using the *ensemble* of samples for all time-series instances i but with the same n, and the PDF is then somewhat

local in time. Gómez-Herrero et al. [26] use a hybrid ensemble – sliding-window approach, estimating PDFs over values $x_{i,n}$ for all trials i within some time-window $t-\sigma \leq n \leq t+\sigma$, giving the measures a local flavour (discussed further in the chapter by Vicente in this book). Also, note that TRENTOOL (transfer entropy toolbox) [49] implements such an ensemble approach for PDF estimation. For *ergodic* processes, the time-window and ensemble approaches are theoretically equivalent.

Now, note that the sliding-window technique described above only refers to constructing the PDF using all observations from that window – it does not force us to compute the average measure, e.g. $H(X)$, over all observations in that window $\{x_{n-T},\ldots,x_{n+T}\}$. Instead, once the PDF is obtained, we may evaluate the local values of entropy and (conditional) mutual information. Averaging can of course be done, e.g. [90], but while averaging in a sliding-window approach does provide a more local measure than averaging over all available observations in the time series X, it is not local in the same sense as the term is used here (i.e. it does not look at the information involved in a computation at a *single specific time step*).

Still on averages, recall that average information-theoretic measures represent averages over local measures at each observation (see (21)). For time-series X, if the whole series is stationary (or if we look at data from identical phases of a cyclostationary process) then we can take the time-average of all local values in order to compute the relevant averaged information-theoretic measure, i.e.:

$$H(X) = \langle h(x_n) \rangle_n \,. \tag{28}$$

Alternatively, if we are taking an *ensemble approach* with observations $x_{i,n}$ for each time series realisation or trial X_i, then we can take an average across all realisations, e.g.:

$$H(X_n) = \langle h(x_{i,n}) \rangle_i \,, \tag{29}$$

to compute an average measure at the given time index n (across realisations or trials). Indeed, this approach can be quite useful to obtain a "local" quantity in time $H(X_n)$, while mitigating against the large variance in local values (noted in [26]). Of course, the PDFs could be estimated using a hybrid ensemble – sliding-window approach, as noted above [26].

3.4 Estimating the local quantities

As described above, appropriately selecting the observations to use in the PDF is one challenge associated with estimating these local quantities properly. Another challenge is to select the type of estimator to use, and to properly extract local probability estimates from it for evaluating the local information quantities. Full details on information-theoretic estimators are given in a separate chapter of this book by

Vicente. In this subsection we specifically describe evaluation of the *local* quantities using various estimators. [4]

When we have *discrete-valued* data, estimating the local measures is relatively straightforward. One simply counts the matching configurations in the available data to obtain the relevant probability estimates ($\hat{p}(x \mid y)$ and $\hat{p}(x)$ for mutual information), and then uses these values directly in the equation for the given local quantity (e.g. (25) for local mutual information) as a plug-in estimate.

For *continuous-valued* data where we deal with the differential entropy [13] and probability density functions, estimation of the local quantities is slightly more complicated and depends on the estimator being used.

Using *kernel-estimators* (e.g. see [82, 41]), the relevant probabilities (e.g. $\hat{p}(x \mid y)$ and $\hat{p}(x)$ for mutual information) are estimated with kernel functions, and then these values are used directly in the equation for the given local quantity (e.g. (25)) as a plug-in estimate (see e.g. [61]).

With the improvements to kernel-estimation for mutual information suggested by *Kraskov et al.* [45, 44] (and extended to conditional mutual information and transfer entropy by [24, 26]), the PDF evaluations are effectively bypassed, and for the average measure one goes directly to estimates based on nearest neighbour counts n_x and n_y in the marginal spaces for each observation. For example, for Kraskov's algorithm 1 we have:

$$I(X;Y) = \psi(k) - \left\langle \psi(n_x+1) + \psi(n_y+1) \right\rangle + \psi(N), \tag{30}$$

where ψ denotes the digamma function, and the values are returned in nats rather than bits. Local values can be extracted here simply by unrolling the expectation values and computing the nearest neighbour counts only at the given observation (x,y), e.g. for algorithm 1:

$$i(x;y) = \psi(k) - \psi(n_x+1) - \psi(n_y+1) + \psi(N). \tag{31}$$

This has been observed as a "time-varying estimator" in [26] and used to estimate the local transfer entropy in [50] and [89].

Using *permutation entropy* approaches [3] (e.g. symbolic transfer entropy [87]), the relevant probabilities are estimated based on the relative ordinal structure of the joint vectors, and these values are directly used in the equations for the given quantities as plug-in estimates (e.g. see local symbolic transfer entropy in [72]).

Finally, using a *multivariate Gaussian model* for $\mathbf{X}$ (which is of d dimensions), the average entropy has the form [13]:

$$H(\mathbf{X}) = \frac{1}{2} \ln \left((2\pi e)^d \mid \Omega \mid \right), \tag{32}$$

(in nats) where $\mid \Omega \mid$ is the determinant of the $d \times d$ covariance matrix $\Omega = \overline{\mathbf{X}^T \mathbf{X}}$ (for row vectors $\mathbf{X}$), and the overbar "represents an average over the statistical en-

[4] Open-source code is available for local information-theoretic measures (using all of the estimator types considered here) in the *Java Information Dynamics Toolkit* on Google code [51].

semble" [6]. Any standard information-theoretic measure of the variables (at the same time step), e.g. mutual information, can then be obtained from sums and differences of these joint entropies. While the PDFs were again effectively bypassed in the average, the local entropies (and by sums and difference other local measures) can be obtained by first reconstructing the probability of a given observation $\mathbf{x}$ in a multivariate process with covariance matrix Ω :

$$p(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^d \mid \Omega \mid^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)\Omega^{-1}(\mathbf{x}-\mu)^T\right), \tag{33}$$

(where μ is the expectation value of $\mathbf{x}$), then using these values directly in the equation for the given local quantity as a plug-in estimate.[5]

4 Local measures of information processing

In this section, we build on the fundamental quantities of information theory, our first look at dynamic measures of information, and on the dynamics of local information measures in time, to present measures of the dynamics of information processing. We briefly review the framework for information dynamics which was recently introduced in [58, 59, 60, 62, 52].

The fundamental question the measures of this framework address is: *"where does the information in a random variable X_{n+1} in a time series come from?"*. This question is addressed in terms of information from the past of process X (i.e. the information *storage*), information contributed from other source processes Y (i.e. the information *transfer*), and how these sources combine (information *modification*). Here we describe local measures of information storage and transfer, and refer the reader to [60, 23, 53] regarding information modification.

4.1 Local information storage

The **active information storage** A_X was introduced [62] to measure how much of the information from the past of the process is observed to be *in use* in computing its next state.[6] The active information storage A_X is the average mutual information between realizations $\mathbf{x}_n^{(k)}$ of the past state $\mathbf{X}_n^{(k)}$ (as $k \to \infty$) and the corresponding

[5] See the next section, Sect. 4.2, regarding how this method can be used to produce a local Granger causality, as a local transfer entropy using a Gaussian model estimator.

[6] This contrasts with related measures including: the *statistical complexity* [15] which measures *all* information stored by the system which *may be used* in the future; and the *excess entropy* [31, 14] which measures that information which *is used* by the system *at some point* in the future. Of course, this means that the excess entropy measures information storage that will possibly but not necessarily be used at the next time step $n+1$, which is greater than or equal to that measured by the active information storage. See further discussion in [62].

realizations x_{n+1} of the *next value* X_{n+1} of a given time series process X:

$$A_X = \lim_{k\to\infty} A_X(k), \tag{34}$$

$$A_X(k) = I\left[\mathbf{X}_n^{(k)}; X_{n+1}\right]. \tag{35}$$

We note that the limit $k \to \infty$ is required in general so as to capture all relevant information in the past of X, unless the next value x_{n+1} is conditionally independent of the far past values $x_{n-k}^{(\infty)}$ given $x_n^{(k)}$ [62]. Empirically of course, one is limited to finite-k estimates $A_X(k)$.

Now, the **local active information storage** $a_X(n+1)$ is the local mutual information between realizations $\mathbf{x}_n^{(k)}$ of the past state $\mathbf{X}_n^{(k)}$ (as $k \to \infty$) and the corresponding realizations x_{n+1} of the *next value* X_{n+1}. This is computed as described for local mutual information values in Sect. 3.2. The average active information storage A_X is the expectation of these local values:

$$A_X = \langle a_X(n+1) \rangle, \tag{36}$$

$$a_X(n+1) = \lim_{k\to\infty} a_X(n+1,k), \tag{37}$$

$$A_X(k) = \langle a_X(n+1,k) \rangle, \tag{38}$$

$$a_X(n+1,k) = i(\mathbf{x}_n^{(k)}; x_{n+1}), \tag{39}$$

$$= \log_2 \frac{p(x_{n+1} \mid \mathbf{x}_n^{(k)})}{p(x_{n+1})}. \tag{40}$$

The local values of active information storage measure the dynamics of information storage at different time points within a system, revealing to us how the use of memory fluctuates during a process. Where the observations used for the relevant PDFs are from the whole time series of a process (under an assumption of stationarity, as outlined in Sect. 3.3), then the average $A_X(k)$ is the time-average of the local values $a_X(n+1,k)$.

We also note that since [62]:

$$A(X) = H(X) - H_\mu(X), \tag{41}$$

then the limit in (34) exists for stationary processes (i.e. $A(X)$ converges with $k \to \infty$). A proof for convergence of $a(x_{n+1})$ with $k \to \infty$ remains a topic for future work.

As described for the local mutual information in Sect. 3.2, $a_X(n+1)$ may be positive or negative, meaning the past history of the process can either positively inform us or actually *misinform* us about its next value [62]. An observer of the process is misinformed where, conditioned on the past history the observed outcome was *relatively* unlikely as compared to the unconditioned probability of that outcome (i.e. $p(x_{n+1} \mid \mathbf{x}_n^{(k)}) < p(x_{n+1})$). In deterministic systems (e.g. CAs), negative local active information storage means that there must be strong information transfer from other causal sources.

4.2 Local information transfer

Information transfer is defined as the amount of information that a source process provides about a target (or destination) process' next state that was not contained in the target's past. This definition pertains to Schreiber's **transfer entropy** measure [82], which has become a very popular tool in complex systems in general (e.g. [96, 64, 73, 5, 59, 55, 7]) and in computational neuroscience in particular (e.g. [91, 49, 40, 88, 54, 19]).

The transfer entropy (TE) [82] captures the average mutual information from realizations $\mathbf{y}_n^{(l)}$ of the state $\mathbf{Y}_n^{(l)}$ of a source time-series process Y to the corresponding realizations x_{n+1} of the next value X_{n+1} of the target time-series process X, conditioned on realizations $\mathbf{x}_n^{(k)}$ of the previous state $\mathbf{X}_n^{(k)}$:

$$T_{Y\to X}(l) = \lim_{k\to\infty} T_{Y\to X}(k,l), \tag{42}$$

$$T_{Y\to X}(k,l) = I\left[\mathbf{Y}_n^{(l)}; X_{n+1} \mid \mathbf{X}_n^{(k)}\right]. \tag{43}$$

Schreiber emphasized that, unlike the (unconditioned) time-differenced mutual information, the transfer entropy was a properly directed, dynamic measure of information transfer rather than shared information.

There are a number of important considerations regarding the use of this measure. These are described more fully in the chapter by Wibral in this book, and summarised as follows.

First, in general, one should take the limit as $k \to \infty$ in order to properly embed or represent the previous state $\mathbf{X}_n^{(k)}$ as relevant to the relationship between the next value X_{n+1} and the source $\mathbf{Y}_n^{(l)}$ [59]. Note that k can be limited here where the next value x_{n+1} is conditionally independent of the far past values $x_{n-k}^{(\infty)}$ given $(x_n^{(k)}, y_n)$. We observe that this historical information conditioned on by the transfer entropy is exactly that provided by the active information storage. As such, setting k properly in this manner gives the observer the perspective to properly separate information storage and transfer in the distributed computation in the systems, and allows one to interpret the transfer entropy as properly representing information transfer [59, 56]. Empirically of course one is restricted to finite-k estimates $T_{Y\to X}(k,l)$.

Also, note that the transfer entropy can be defined for an arbitrary source-target delay, i.e. measuring $I\left[\mathbf{Y}_{n-u}^{(l)}; X_{n+1} \mid \mathbf{X}_n^{(k)}\right]$, and indeed that this should be done for the appropriate causal delay $u > 0$ [93]. For ease of presentation here, we describe the measures for $u = 1$ only, though all are straightforward to generalise.

Furthermore, considering the source *state* $\mathbf{y}_n^{(l)}$ rather than a scalar y_n is most appropriate where the observations y mask a hidden Markov process which is causal to X, or where multiple past values of Y in addition to y_n are causal to x_{n+1}. Otherwise, where y_n is directly causal to x_{n+1}, and where it is the only direct causal source in Y, we use only $l = 1$ [59, 56].

Finally, for proper interpretation as information transfer, Y is constrained among the causal information contributors to X [56]. We have also provided a thermodynamic interpretation of transfer entropy in [79], as being proportional to external entropy production, possibly due to irreversibility.

Now, we continue on to extract the **local transfer entropy** $t_{Y\to X}(n+1)$ [59] as a local conditional mutual information using the approach described in Sect. 3.2. It is the amount of information transfer attributed to the specific configuration or realization $(x_{n+1}, \mathbf{x}_n^{(k)}, \mathbf{y}_n^{(l)})$ at time step $n+1$; i.e. the amount of information transfered from process Y to X at time step $n+1$:

$$T_{Y\to X}(l) = \langle t_{Y\to X}(n+1,l)\rangle, \tag{44}$$

$$t_{Y\to X}(n+1,l) = \lim_{k\to\infty} t_{Y\to X}(n+1,k,l), \tag{45}$$

$$T_{Y\to X}(k,l) = \langle t_{Y\to X}(n+1,k,l)\rangle, \tag{46}$$

$$t_{Y\to X}(n+1,k,l) = i(\mathbf{y}_n^{(l)}; x_{n+1} \mid \mathbf{x}_n^{(k)}), \tag{47}$$

$$= \log_2 \frac{p(x_{n+1} \mid \mathbf{x}_n^{(k)}, \mathbf{y}_n^{(l)})}{p(x_{n+1} \mid \mathbf{x}_n^{(k)})}. \tag{48}$$

These local information transfer values measure the dynamics of transfer in time between any given pair of processes within a system, revealing to us how information is transferred across the system in time and space. Fig. 4.2 indicates a local transfer entropy measurement for a pair of processes $Y \to X$.

As above, where the observations used for the relevant PDFs are from the whole time series of the processes (under an assumption of stationarity, as outlined in Sect. 3.3) then the average $T_{Y\to X}(k,l)$ is the time-average of the local transfer values $t_{Y\to X}(n+1,k,l)$.

As described for the local conditional mutual information in Sect. 3.2, $t_{Y\to X}(n+1)$ may be positive or negative, meaning the source process can either positively inform us or actually *misinform* us about the next value of the target (in the context of the target's past state) [59]. An observer of the process is misinformed where, conditioned on the source and the past of the target the observed outcome was *relatively* unlikely, as compared to the probability of that outcome conditioning on the past history only (i.e. $p(x_{n+1} \mid \mathbf{x}_n^{(k)}, \mathbf{y}_n^{(l)}) < p(x_{n+1} \mid \mathbf{x}_n^{(k)})$).

Noting the equivalence of the transfer entropy and the concept of Granger causality [28] when the transfer entropy is estimated using a Gaussian model [4], we observe that the local transfer entropy – when estimated with a Gaussian model as described in Sect. 3.4 – directly gives a *local Granger causality* measurement .

Now, the transfer entropy may also be conditioned on other possible sources Z to account for their effects on the target. The **conditional transfer entropy** was introduced for this purpose [59, 60]:

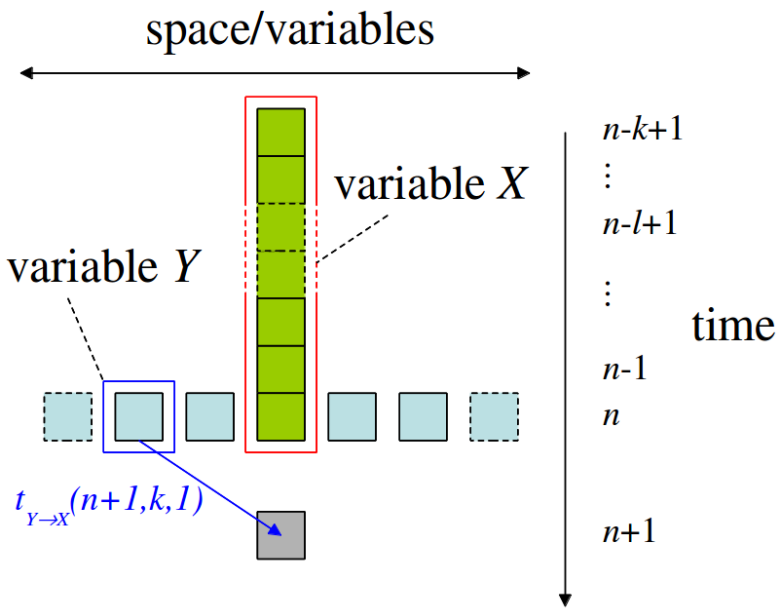


Fig. 1 Local transfer entropy $t_{Y\to X}(n+1,k,l=1)$ indicated by the blue arrow: information contained in the realization y_n of the source variable Y about the next value x_{n+1} of the destination variable X at time $n+1$, in the context of the corresponding realization $x_n^{(k)}$ of the destination's past state.

$$T_{Y\to X|Z}(l) = \lim_{k\to\infty} T_{Y\to X|Z}(k,l), \tag{49}$$

$$T_{Y\to X|Z}(k,l) = I\left[\mathbf{Y}_n^{(l)}; X_{n+1} \mid \mathbf{X}_n^{(k)}, Z\right], \tag{50}$$

Note that Z may represent an embedded state of another variable and/or be explicitly multivariate. Transfer entropies conditioned on other variables have been used in several biophysical and neuroscience applications, e.g. [20, 21, 88].

We also have the corresponding **local conditional transfer entropy**:

$$T_{Y\to X|Z}(k,l) = \left\langle t_{Y\to X|Z}(n+1,k,l)\right\rangle, \tag{51}$$

$$t_{Y\to X|Z}(n+1,k,l) = \log_2 \frac{p(x_{n+1} \mid \mathbf{x}_n^{(k)}, \mathbf{y}_n^{(l)}, z_n)}{p(x_{n+1} \mid \mathbf{x}_n^{(k)}, z_n)}, \tag{52}$$

$$= i(\mathbf{y}_n^{(l)}; x_{n+1} \mid \mathbf{x}_n^{(k)}, z_n). \tag{53}$$

Of course, this extra conditioning can prevent the (redundant) influence of a common drive Z from being attributed to Y, and can also include the synergistic contribution when the source Y acts in conjunction with another source Z (e.g. where X is the outcome of an XOR operation on Y and Z).

We specifically refer to the conditional transfer entropy as the **complete transfer entropy** (with notation $T^c_{Y\to X}(k,l)$ and $t^c_{Y\to X}(n+1,k,l)$ for example) when it conditions on all other causal sources Z to the target X [59]. To differentiate the conditional and complete transfer entropies from the original measure, we often refer to $T_{Y\to X}$ simply as the *apparent* transfer entropy [59] - this nomenclature conveys that the result is the information transfer that is apparent without accounting for other sources.

Finally, note that one can decompose the mutual information from a set of sources to a target as a sum of incrementally conditioned mutual information terms [60, 56, 53]. For example, for a two source system we have:

$$\begin{aligned} I(X_{n+1};\{\mathbf{X}_n^{(k)},Y_{1,n},Y_{2,n}\}) &= I(X_{n+1};\mathbf{X}_n^{(k)}) + I(X_{n+1};Y_{1,n} \mid \mathbf{X}_n^{(k)}) + \\ &\quad + I(X_{n+1};Y_{2,n} \mid \mathbf{X}_n^{(k)},Y_{1,n}), \\ &= A_X(k) + T_{Y_1\to X}(k) + T_{Y_2\to X|Y_1}(k). \end{aligned} \tag{54}$$

This equation could be reversed in the order of Y_1 and Y_2, and its correctness is independent of k (so long as k is large enough to capture the causal sources in the past of the target). Crucially, this equation reveals the nature in which information storage (A_X) and transfer ($T_{Y_1\to X}$, etc.) are complementary operations in distributed computation.

5 Local information processing in cellular automata

In this section, we review the application of local information storage and transfer measures to cellular automata (as first presented in [58, 59, 56, 60, 62, 61]), in order to demonstrate the ability of the local measures to reveal deeper insights into the dynamics of complex systems than their averaged and more well-known counterparts.

Cellular automata (CAs) are discrete dynamical systems with an array of cells that synchronously update their value as a function of a fixed number of spatial neighbours cells using a uniform rule [97]. The update rule is specified by listing the next value for a given cell as a function of each possible configuration of its neighbourhood in a rule table – see Table 1 – and summarising this specification in a single number (known as a Wolfram number; see [97]). We focus here on Elementary CAs (ECAs), which are 1D arrays of binary-valued cells with one neighbour on either side.

Although the behaviour of each individual cell in a CA is very simple, the (non-linear) interactions between all cells can lead to very intricate global behaviour, meaning CAs have become a classic example of self-organised complex dynamics. Of particular importance, CAs have been used to model real-world spatial dynamical processes, including fluid flow, earthquakes and biological pattern formation [70]. Indeed, CAs have even been used in neural network models to study criticality in avalanches of activity [75, 67]. While they may not be the most realistic microscopic

Table 1 Rule table for ECA rule 54. The Wolfram rule number for this rule table is composed by taking the next cell value for each configuration, concatenating them into a binary code starting from the bottom of the rule table as the most significant bit (e.g. b00110110 here), and then forming the decimal rule number from that binary encoding.

Neighbourhood configuration for cell i at time n			Next cell value $x_{i,n+1}$ at time $n+1$
cell $x_{i-1,n}$ value (left)	cell $x_{i,n}$ value	cell $x_{i+1,n}$ value (right)	
0	0	0	**0**
0	0	1	**1**
0	1	0	**1**
0	1	1	**0**
1	0	0	**1**
1	0	1	**1**
1	1	0	**0**
1	1	1	**0**

neural model available, it is certainly true that CAs can exhibit certain phenomena that are of particular interest in neuroscience, including avalanche behaviour (e.g. [75, 80, 47, 67]) and coherent propagating wave-like structures (e.g. [27, 17]).

Indeed, the presence of such coherent emergent structures: *particles*, *gliders*, *blinkers* and *domains*; is what has made CAs so interesting in complex systems science in general. A domain is a set of background configurations in a CA, any of which will update to another configuration in the set in the absence of any disturbance. Domains are formally defined by computational mechanics as spatial process languages in the CA [33]. Particles are considered to be dynamic elements of coherent spatiotemporal structure, which are disturbances or lie in contrast to the background domain. Gliders are regular particles, blinkers are stationary gliders. Formally, particles are defined by computational mechanics as a boundary between two domains [33]; as such, they can be referred to as *domain walls*, though this term is usually reserved for irregular particles. Several techniques exist to *filter* particles from background domains (e.g. [29, 30, 33, 34, 98, 36, 37, 84, 59, 60, 62]).

These emergent structures have been quite important to studies of distributed computation in CAs, for example in the design or identification of universal computation (see [70]), and analyses of the dynamics of intrinsic or other specific computation ([46, 33, 71]). This is because these studies typically discuss the computation in terms of the three primitive functions of computation and their apparent analogues in CA dynamics [70, 46]:

- blinkers as the basis of information storage, since they periodically repeat at a fixed location;
- particles as the basis of information transfer, since they communicate information about the dynamics of one spatial part of the CA to another part; and
- collisions between these structures as information modification, since collision events combine and modify the local dynamical structures.

Previous to the work reviewed here however, these analogies remained conjecture only, based on qualitative observation of CA dynamics. In the following subsections,

we review the applications [59, 60, 62, 58, 56] of the local information storage and transfer measures described in Sect. 4 to cellular automata.

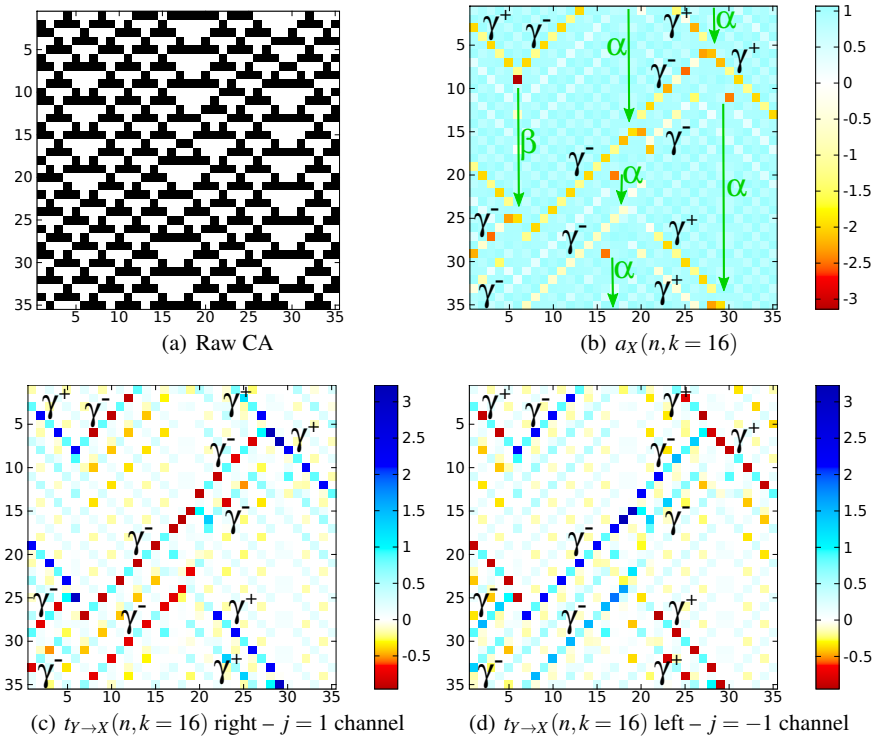
These experiments involved constructing 10 000 cell 1-dimensional CAs, and executing the relevant update rules to generate 600 time steps of dynamics. All resulting 6×10^6 observations of cell-updates are then used to compose the relevant PDFs, and the local measures of information storage and transfer were computed for each observation using these PDFs. Specifically, local active information storage $a_X(n, k = 16)$ is computed for each cell X for each time step n, while local transfer entropy $t_{Y \to X}(n, k = 16, l = 1)$ is computed for each time step n for each target cell X and for the two causal sources Y on either side of X (referred to as channels $j = 1$ and -1 for transfer across 1 cell to the right or left). The use of all observations across all cells and time steps implies an assumption of stationarity here. This is justified in that the large CA length and relativity short number of time steps (and ignoring of initial steps) is designed to ensure that an attractor is not reached while the typical transient dynamics of the CA are well-sampled. Note also that $l = 1$ is used since we directly observe the interacting values and only one previous time step is a causal source here. As such, in line with (54) we have

$$I(X_{n+1}; \{\mathbf{X}_n^{(k)}, Y_{l,n}, Y_{r,n}\}) = A_X(k) + T_{Y_l \to X}(k) + T_{Y_r \to X | Y_l}(k), \tag{55}$$

where Y_l represents the causal source to the left (channel $j = 1$) and Y_r the causal source to the right (channel $j = -1$) – although their placement is interchangeable in this equation.

Sample results of this application are displayed for rules 54 and 18 in Fig. 2 and Fig. 3. The figures displayed here were produced using the open source *Java Information Dynamics Toolkit* (JIDT) [51], which can be used in Matlab, Octave and Python as well as Java. All results can be reproduced using the Matlab/Octave script `DirectedMeasuresChapterDemo2013.m` in the `demos/octave/-CellularAutomata` example distributed with this toolkit.

These applications provided the first quantitative evidence for the above conjectures, and are discussed in the following subsections. But the most important result for our purposes is that **the local measures reveal richly-structured spatiotemporal profiles of the information storage and transfer dynamics here**, with interesting local features revealed at various points in space-time. It is simply not possible for these dynamics to be revealed by the average measures, be they averages across all cells and times or averages just across all cells in time. These features are uniquely provided by considering the *local* dynamics of information processing in CAs, and are discussed in the following subsections.



(a) Raw CA (b) $a_X(n,k=16)$

(c) $t_{Y\to X}(n,k=16)$ right – $j=1$ channel (d) $t_{Y\to X}(n,k=16)$ left – $j=-1$ channel

Fig. 2 Local information dynamics in ECA **rule 54** for the raw values in (a) (black for "1", white for "0"). 35 time steps are displayed for 35 cells, and time increases down the page for all CA plots. All units are in bits. (b) Local active information storage; Local apparent transfer entropy: (c) one cell to the right, and (d) one cell to the left per time step.

5.1 Blinkers and background domains as information storage entities

The first and most expected result is that **blinkers (regular, stationary particles) and regular background domains are dominant information storage entities** [62], e.g. see Fig. 2(b). This is because these structures are temporally periodic, and so the past state of a cell $\mathbf{x}_n^{(k)}$ is highly predictive of the next value x_{n+1} – this means that we have $p(x_{n+1} \mid \mathbf{x}_n^{(k)}) > p(x_{n+1})$, giving large positive values of $a_X(n+1,k)$ via (40).

In contrast, we see in Fig. 2(b) and Fig. 3(b) that moving particle structures (both regular gliders and domain walls) are associated with negative local information storage $a_X(n+1,k)$. This is because at these locations, the past state of a cell $\mathbf{x}_n^{(k)}$ is part of the background domain and observing it would normally predict that the background domain continues. Since a particle is encountered at the cell instead

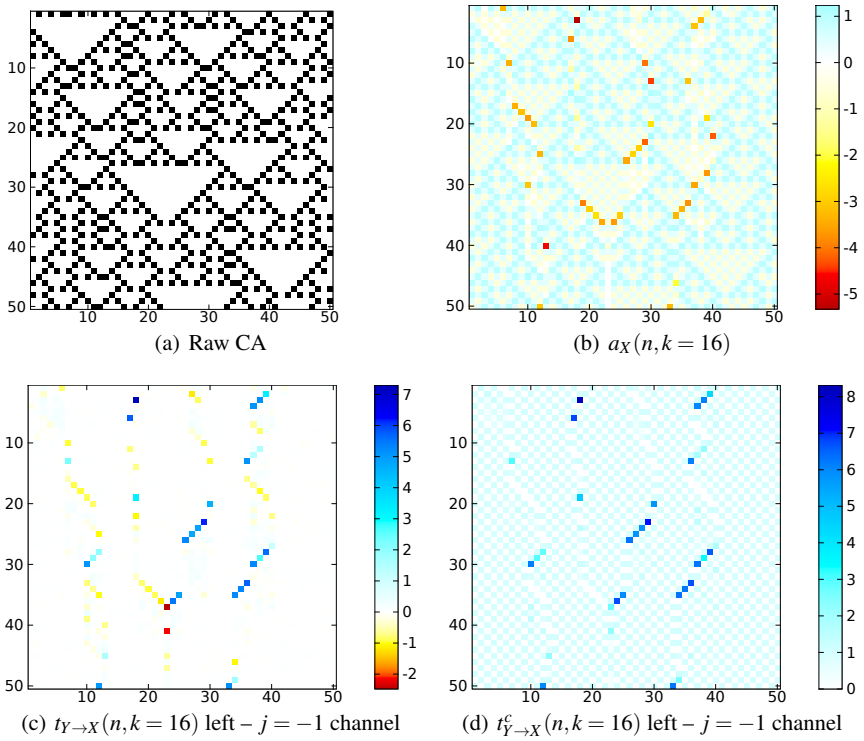


Fig. 3 Local information dynamics in ECA **rule 18** for the raw values in (a) (black for "1", white for "0"). 50 time steps are displayed for 50 cells, and all units are in bits. (b) Local active information storage; (c) Local apparent transfer entropy one cell to the left per time step; (d) Local complete transfer entropy one cell to the left per time step.

however, this past state $\mathbf{x}_n^{(k)}$ is in fact misinformative about the next value x_{n+1}. That is to say, we have $p(x_{n+1} \mid \mathbf{x}_n^{(k)}) < p(x_{n+1})$, giving negative values of $a_X(n+1,k)$ via (40). We note that these misinformative values can only occur (for this deterministic system) where another information source is having a relatively large predictive effect on the target – to explore these further, we turn our attention to local information transfer in the next subsection.

Finally, we note that these results required a large enough k to properly capture the past state of the cell, and could not be observed with a value say of $k = 1$ (as discussed in [62]).

5.2 Particles, gliders and domain walls as dominant information transfer entities

Perhaps the most important result from our application to CAs is that local information transfer is typically strongly positive at moving particles in comparison to blinkers and background domains [59]. To clarify, this is when the local information transfer is measured at a particle in the same direction or channel j as the macroscopic motion of that particle. For example, see the highlighting of left and right moving gliders for rule 54 in Fig. 2(c) and Fig. 2(d) by transfer entropy to the left and right respectively, and similarly for the left moving sections of domain walls for rule 18 in Fig. 3(c) and Fig. 3(d) by transfer entropy to the left (TE to right omitted). In these examples, the past state of the target cell $\mathbf{x}_n^{(k)}$ is part of the background domain and so is misinformative about the next value x_{n+1} where the particle is encountered. In contrast, the source cell y_n which is in the particle at the previous time step n (be that the left or right neighbour, as relevant for that particular particle) is highly predictive about the next value of the target (in the context of its past). As such, we have $p(x_{n+1} \mid \mathbf{x}_n^{(k)}, y_n) > p(x_{n+1} \mid \mathbf{x}_n^{(k)})$, giving large positive values of $t_{Y \to X}(n+1,k)$ via (48).

These results for local transfer entropy are particularly important because they provided the **first quantitative evidence for the long-held conjecture that particles are the dominant information transfer agents in CAs**. As stated above, it is simply not possible for these space-time specific dynamics to be revealed by the average transfer entropy, it specifically requires the local transfer entropy. Furthermore, the average values do not give so much as a hint towards the complexities of these local dynamics: ECA rule 22 has much larger average transfer entropy values than rule 54 (0.19 versus 0.08 bits for each, respectively, in both left and right directions), yet has no emergent self-organized particle structures [61].

As per the information storage results, we note that these results required a large enough k to properly capture the past state of the cell, and could not be observed with a value say of $k = 1$ (as discussed in [59]). When linked to the result of misinformative storage at the particles from Sect. 5.1, we see again the complementary nature of information storage and transfer.

It is important to note that particles are not the only points with positive local transfer entropy. Small positive non-zero values are also often measured in the domain and in the orthogonal direction to glider motion in space-time (e.g. see Fig. 2(d)) [59]. These correctly indicate non-trivial information transfer in these regions (e.g. indicating the *absence* of a glider), though they are dominated by the positive transfer in the direction of glider motion.

5.3 Sources can be locally misinformative

Next, we note that local information transfer is often found to be negative at moving particles, when measured in the orthogonal direction to macroscopic particle motion in space-time [59]. For example, see the right-moving gliders in Fig. 2(d) or right-moving domain walls in Fig. 3(c)). This is because the source Y here, being on the opposite side of the target to the incoming particle and therefore still part of the domain observed in the target's past, would suggest that this domain pattern would continue, which is misinformative. That is to say, we have here $p(x_{n+1} \mid \mathbf{x}_n^{(k)}, y_n) < p(x_{n+1} \mid \mathbf{x}_n^{(k)})$, giving negative values of $t_{Y \to X}(n+1,k)$ via (48).

As described in Sect. 4.2, a source can be locally misinformative but must be positively informative on average (or at least provide zero information). These negative or misinformative values are quite useful, since they imply that there is an extra feature in the dynamics that is unaccounted for in the past of the source and target alone. In the case of deterministic systems, this means that more sources must be examined to explain the dynamics, as explored in the next subsection.

5.4 Conditional transfer entropy is complementary

Fig. 3(d) displays a profile of the local conditional transfer entropy $t_{Y \to X|Z}$ applied to rule 18 (discussed in detail in [59]). This is the transfer entropy from the source cell Y on the right of the target X, conditioned on the other source cell Z on the left. Because we condition on all of the other causal sources here, this measurement may also be referred to as a complete transfer entropy [59].

This profile is rather different to that of the apparent transfer entropy $t_{Y \to X}$ for the same channel (i.e. from the same relative source) displayed in Fig. 3(c). The first noticeable difference is the checkerboard pattern of transfer in the background domain, which is only visible with the conditional measure. This pattern forms due to complex dynamics in the domain here, with two interleaving phases. The first phase occurs at every second cell (both in space and time), and is simply a '0' – at these cells there is strong information storage alone (see Fig. 3(b)) because the cell value is predictable from its past (which predicts the phase accurately). The other phase occurs at the alternate cells, and is a '0' or a '1' as determined via an exclusive OR (or XOR) operation between the neighbouring left and right cells. As such, apparent transfer entropy from either left or right cell alone provides almost no information about the next value (hence absence of apparent transfer in the domain – see Fig. 3(c)), whilst conditional transfer entropy provides full information about the next value because the other contributing cell is taken into account (hence the strong conditional transfer at every second cell in Fig. 3(d)).

The other noticeable difference between these profiles is that the conditional transfer entropy does not have any negative local values, unlike the apparent transfer entropy. This is because examining the source in the context of all other causal

sources in this deterministic system necessarily provides more information than not examining the source. That is to say, there are no unaccounted sources here which could mislead the observer, unlike that possibility for the apparent transfer entropy.

There are two key messages from the comparison of these measures:

1. **The apparent and conditional transfer entropy reveal different aspects of the dynamics of a system** – neither is more correct than the other; they are both useful and complementary. This is a particularly important message, since often the importance of conditioning "out" all other sources using a conditional measure is emphasised, without acknowledging the complementary utility retained by the pairwise transfer entropy. Both are required to have a full picture of the dynamics of a system;
2. The differences in local dynamics that they reveal simply cannot be observed here by using the average of each measure alone.

5.5 Contrasting information transfer and causal effect

Finally, we note that differences between the concepts of information transfer (as captured by the transfer entropy) and causal effect are now well established [2, 56, 11]. We briefly review how the local perspective of transfer entropy was used to provide insight into these differences in [56].

Causal effect refers to the extent to which the source variable has a direct influence or drive on the next state of a target variable, i.e. "if I change the value of the source, to what extent does that alter the value of the target?" [74, 2, 56]. In this light, consider the causal effect of the left cell $x_{i-1,n}$ in the seventh row of the rule table for rule 54 in Table 1, i.e. "1 1 0 $\rightarrow$ 0". Altering the value of this source has a clear causal effect on the target, since it changes the rule being executed to "0 1 0 $\rightarrow$ 1" (i.e. we have a different outcome at the target). Crucially though, this particular configuration ("1 1 0 $\rightarrow$ 0") is observed both in the (right-moving) gliders and in the background domain of rule 54. This means that the same causal effect occurs in both types of dynamics.[7]

This is quite different to our interpretation of information transfer in the previous sections however. This interpretation can be restated as: predictive information transfer refers to the amount of information that a source variable adds to the state change of a target variable; i.e. "if I know the state of the source, how much does that help to predict the state change of the target?" [56]. In dealing with *state* updates of the target, and in particular in separating information storage from transfer, the transfer entropy has a very different perspective to causal effect. As we have seen, local transfer entropy attributes large positive local values at the gliders here, because the source cells help prediction in the context of a target's past, but attributes

[7] Reference [56], which covers this issue in more depth, explores measuring the causal effect in these dynamics using the measure presented in [2].

vanishing amounts in the domain, where stored information from a target's past is generally sufficient for prediction.

Again, neither perspective is more correct than the other – they both provide useful insights and are complementary. This argument is explored in more depth in [56]. Crucially, these insights are only fully revealed with our *local* perspective of information dynamics here.

6 Discussion: relevance of local measures to computational neuroscience

In the previous section, we have demonstrated that local transfer entropy and the associated measures of local information dynamics provide key insights into local information processing in cellular automata that cannot be provided with traditional average information-theoretic measures. We have gone on to use these local techniques to provide similar insights in other systems, such as:

- visualising coherent waves of motion in flocks (or swarms) as information cascades spreading across the flock (as previously conjectured, [12]) using local transfer entropy [92];
- revealing coherent information transfer waves in modular robots [57];
- demonstrating information transfer as a key driver in the dynamics of network synchronization processes, with local values dropping to zero (i.e. the synchronized state has been "computed") before it is otherwise apparent that a synchronized state has been either reached or determined [9].

We can reasonably expect local information transfer and storage to provide new insights in a computational neuroscience setting also. As described earlier, avalanche behaviour (e.g. [80, 47, 75]) and coherent propagating wave-like structures (e.g. [27, 17]) are of particular interest in neuroscience, and particles and gliders bear more than a passing resemblance to these coherent structures. Given that local transfer entropy has been used to provide the first quantitative evidence that similar propagating coherent structures in other domains are information transfer entities (e.g. particles and gliders in cellular automata [59], above, motion in flocks and swarms [92], and in modular robotics [57]), one expects that this measure will be used to provide similar insights into these structures in neural systems.

Yet local transfer entropy will find much more broad application than simply identifying local coherent structure. It offers the opportunity to answer the question:

> "Precisely *when* and *where* is information transferred between brain regions?"

The *where* is answerable with average transfer entropy, but the *when* is only precisely answerable with a local approach. This is a fundamentally important question for us to have the opportunity to answer, because it will provide insight into the precise dynamics of how information is stored, transferred and modified in the brain during neural computation.

For example, we have conducted a preliminary study applying this method to a set of fMRI measurements where we could expect to see differences in local information transfer between two conditions at specific time steps [50]. The fMRI data set analyzed (from [86]) is a 'Libet'-style experiment, which contains brain activity recorded while subjects were asked to freely decide whether to push one of two buttons (with left or right index finger). Significant differences (at the group level) were found in the local transfer entropy between left and right button presses from a single source region (e.g. pre-SMA) into the left and right motor cortices respectively. Furthermore, simple thresholding of these local transfer entropy values provides a statistically significant prediction of which button was pressed.

These results are a strong demonstration that local transfer entropy can usefully provide task-relevant insights into when and where information is transferred between brain regions. Once validation studies have been completed in this domain, we expect that further utility will be found for these local information-theoretic measures in computational neuroscience. There are many studies in this domain which will benefit from the ability to view local information storage, transfer and modification operations on a local scale in space and time in the brain.

Acknowledgements The author wishes to thank Michael Wibral for very helpful comments on a draft paper and discussions on the topic, as well as Mikhail Prokopenko, Daniel Polani, Ben Flecker and Paul Williams for useful discussions on these topics.

References

1. Ash, R.B.: Information Theory. Dover Publishers, Inc., New York, USA (1965)
2. Ay, N., Polani, D.: Information Flows in Causal Networks. Advances in Complex Systems **11**(1), 17–41 (2008)
3. Bandt, C., Pompe, B.: Permutation entropy: A natural complexity measure for time series. Physical Review Letters **88**(17) (2002). DOI 10.1103/physrevlett.88.174102. URL `http://dx.doi.org/10.1103/physrevlett.88.174102`
4. Barnett, L., Barrett, A.B., Seth, A.K.: Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. Physical Review Letters **103**(23), 238,701+ (2009)
5. Barnett, L., Bossomaier, T.: Transfer Entropy as a Log-Likelihood Ratio. Physical Review Letters **109**, 138,105+ (2012)
6. Barnett, L., Buckley, C.L., Bullock, S.: Neural complexity and structural connectivity. Physical Review E **79**(5), 051,914+ (2009)
7. Boedecker, J., Obst, O., Lizier, J.T., Mayer, N.M., Asada, M.: Information processing in echo state networks at the edge of chaos. Theory in Biosciences **131**(3), 205–213 (2012)
8. Bressler, S.L., Tang, W., Sylvester, C.M., Shulman, G.L., Corbetta, M.: Top-Down Control of Human Visual Cortex by Frontal and Parietal Cortex in Anticipatory Visual Spatial Attention. Journal of Neuroscience **28**(40), 10,056–10,061 (2008)
9. Ceguerra, R.V., Lizier, J.T., Zomaya, A.Y.: Information storage and transfer in the synchronization process in locally-connected networks. In: Proceedings of the 2011 IEEE Symposium on Artificial Life (ALIFE), pp. 54–61. IEEE (2011)
10. Chávez, M., Martinerie, J., Le Van Quyen, M.: Statistical assessment of nonlinear causality: application to epileptic EEG signals. Journal of Neuroscience Methods **124**(2), 113–128 (2003)

11. Chicharro, D., Ledberg, A.: When Two Become One: The Limits of Causality Analysis of Brain Dynamics. PLoS ONE **7**(3), e32,466+ (2012)
12. Couzin, I.D., James, R., Croft, D.P., Krause, J.: Social Organization and Information Transfer in Schooling Fishes. In: C. Brown, K.N. Laland, J. Krause (eds.) Fish Cognition and Behavior, Fish and Aquatic Resources, pp. 166–185. Blackwell Publishing (2006)
13. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, New York (1991)
14. Crutchfield, J.P., Feldman, D.P.: Regularities Unseen, Randomness Observed: Levels of Entropy Convergence. Chaos **13**(1), 25–54 (2003)
15. Crutchfield, J.P., Young, K.: Inferring statistical complexity. Physical Review Letters **63**(2), 105–108 (1989)
16. Dasan, J., Ramamohan, T.R., Singh, A., Nott, P.R.: Stress fluctuations in sheared Stokesian suspensions. Physical Review E **66**(2), 021,409 (2002)
17. Derdikman, D., Hildesheim, R., Ahissar, E., Arieli, A., Grinvald, A.: Imaging spatiotemporal dynamics of surround inhibition in the barrels somatosensory cortex. The Journal of Neuroscience **23**(8), 3100–3105 (2003)
18. DeWeese, M.R., Meister, M.: How to measure the information gained from one symbol. Network: Computation in Neural Systems **10**, 325–340 (1999)
19. Effenberger, F.: A primer on information theory, with applications to neuroscience (2013). URL `http://arxiv.org/abs/1304.2333`. arXiv:1304.2333
20. Faes, L., Nollo, G., Porta, A.: Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. Physical Review E **83**, 051,112+ (2011)
21. Faes, L., Nollo, G., Porta, A.: Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series. Computers in Biology and Medicine **42**(3), 290–297 (2012)
22. Fano, R.M.: Transmission of information: a statistical theory of communications. M.I.T. Press, Cambridge, MA, USA (1961)
23. Flecker, B., Alford, W., Beggs, J.M., Williams, P.L., Beer, R.D.: Partial information decomposition as a spatiotemporal filter. Chaos: An Interdisciplinary Journal of Nonlinear Science **21**(3), 037,104+ (2011)
24. Frenzel, S., Pompe, B.: Partial Mutual Information for Coupling Analysis of Multivariate Time Series. Physical Review Letters **99**(20), 204,101+ (2007)
25. Friston, K.J., Harrison, L., Penny, W.: Dynamic causal modelling. NeuroImage **19**(4), 1273–1302 (2003)
26. Gomez-Herrero, G., Wu, W., Rutanen, K., Soriano, M.C., Pipa, G., Vicente, R.: Assessing coupling dynamics from an ensemble of time series (2010). URL `http://arxiv.org/abs/1008.0539`. arXiv:1008.0539
27. Gong, P., van Leeuwen, C.: Distributed Dynamical Computation in Neural Circuits with Propagating Coherent Activity Patterns. PLoS Computational Biology **5**(12) (2009)
28. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**, 424–438 (1969)
29. Grassberger, P.: New mechanism for deterministic diffusion. Physical Review A **28**(6), 3666 (1983)
30. Grassberger, P.: Long-range effects in an elementary cellular automaton. Journal of Statistical Physics **45**(1-2), 27–39 (1986)
31. Grassberger, P.: Toward a quantitative theory of self-generated complexity. International Journal of Theoretical Physics **25**(9), 907–938 (1986)
32. Griffith, V., Koch, C.: Quantifying synergistic mutual information (2012). URL `http://arxiv.org/abs/1205.4265`. arXiv:1205.4265
33. Hanson, J.E., Crutchfield, J.P.: The Attractor-Basin Portait of a Cellular Automaton. Journal of Statistical Physics **66**, 1415–1462 (1992)
34. Hanson, J.E., Crutchfield, J.P.: Computational mechanics of cellular automata: An example. Physica D **103**(1-4), 169–189 (1997)

35. Harder, M., Salge, C., Polani, D.: A Bivariate Measure of Redundant Information (2012). URL http://arxiv.org/abs/1207.2080. arXiv:1207.2080
36. Helvik, T., Lindgren, K., Nordahl, M.G.: Local information in one-dimensional cellular automata. In: P.M.A. Sloot, B. Chopard, A.G. Hoekstra (eds.) Proceedings of the International Conference on Cellular Automata for Research and Industry, Amsterdam, *Lecture Notes in Computer Science*, vol. 3305, pp. 121–130. Springer, Berlin/Heidelberg (2004)
37. Helvik, T., Lindgren, K., Nordahl, M.G.: Continuity of Information Transport in Surjective Cellular Automata. Communications in Mathematical Physics **272**(1), 53–74 (2007)
38. Hinrichs, H., Heinze, H.J., Schoenfeld, M.A.: Causal visual interactions as revealed by an information theoretic measure and fMRI. NeuroImage **31**(3), 1051–1060 (2006)
39. Honey, C.J., Kotter, R., Breakspear, M., Sporns, O.: Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proceedings of the National Academy of Sciences **104**(24), 10,240–10,245 (2007)
40. Ito, S., Hansen, M.E., Heiland, R., Lumsdaine, A., Litke, A.M., Beggs, J.M.: Extending Transfer Entropy Improves Identification of Effective Connectivity in a Spiking Cortical Network Model. PLoS ONE **6**(11), e27,431+ (2011)
41. Kantz, H., Schreiber, T.: Nonlinear Time Series Analysis. Cambridge University Press, Cambridge, MA (1997)
42. Katare, S., West, D.H.: Optimal complex networks spontaneously emerge when information transfer is maximized at least expense: A design perspective. Complexity **11**(4), 26–35 (2006)
43. Kerr, C.C., Van Albada, S.J., Neymotin, S.A., Chadderdon, G.L., Robinson, P.A., Lytton, W.W.: Cortical information flow in parkinson's disease: a composite network/field model. Frontiers in Computational Neuroscience **7**(39) (2013)
44. Kraskov, A.: Synchronization and Interdependence Measures and their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data, *Publication Series of the John von Neumann Institute for Computing*, vol. 24. John von Neumann Institute for Computing, Jülich, Germany (2004)
45. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical Review E **69**(6), 066,138+ (2004)
46. Langton, C.G.: Computation at the edge of chaos: phase transitions and emergent computation. Physica D **42**(1-3), 12–37 (1990)
47. Levina, A., Herrmann, J.M., Geisel, T.: Dynamical synapses causing self-organized criticality in neural networks. Nature Physics **3**(12), 857–860 (2007)
48. Liang, H., Ding, M., Bressler, S.L.: Temporal dynamics of information flow in the cerebral cortex. Neurocomputing **38-40**, 1429–1435 (2001)
49. Lindner, M., Vicente, R., Priesemann, V., Wibral, M.: TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. BMC Neuroscience **12**(1), 119+ (2011)
50. Lizier, J., Heinzle, J., Soon, C., Haynes, J.D., Prokopenko, M.: Spatiotemporal information transfer pattern differences in motor selection. BMC Neuroscience **12**(Suppl 1), P261+ (2011)
51. Lizier, J.T.: JIDT: An information-theoretic toolkit for studying the dynamics of complex systems (2012). URL https://code.google.com/p/information-dynamics-toolkit/
52. Lizier, J.T.: The Local Information Dynamics of Distributed Computation in Complex Systems. Springer Theses. Springer, Berlin / Heidelberg (2013)
53. Lizier, J.T., Flecker, B., Williams, P.L.: Towards a synergy-based approach to measuring information modification. In: Proceedings of the 2013 IEEE Symposium on Artificial Life (ALIFE), pp. 43–51. IEEE (2013)
54. Lizier, J.T., Heinzle, J., Horstmann, A., Haynes, J.D., Prokopenko, M.: Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. Journal of Computational Neuroscience **30**(1), 85–107 (2011)
55. Lizier, J.T., Pritam, S., Prokopenko, M.: Information dynamics in small-world Boolean networks. Artificial Life **17**(4), 293–314 (2011)
56. Lizier, J.T., Prokopenko, M.: Differentiating information transfer and causal effect. European Physical Journal B **73**(4), 605–615 (2010)

57. Lizier, J.T., Prokopenko, M., Tanev, I., Zomaya, A.Y.: Emergence of Glider-like Structures in a Modular Robotic System. In: S. Bullock, J. Noble, R. Watson, M.A. Bedau (eds.) Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems (ALife XI), Winchester, UK, pp. 366–373. MIT Press, Cambridge, MA (2008)
58. Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: Detecting Non-trivial Computation in Complex Dynamics. In: Almeida, L.M. Rocha, E. Costa, I. Harvey, A. Coutinho (eds.) Proceedings of the 9th European Conference on Artificial Life (ECAL 2007), *Lecture Notes in Computer Science*, vol. 4648, pp. 895–904. Springer, Berlin / Heidelberg (2007)
59. Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: Local information transfer as a spatiotemporal filter for complex systems. Physical Review E **77**(2), 026,110+ (2008)
60. Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: Information modification and particle collisions in distributed computation. Chaos **20**(3), 037,109+ (2010)
61. Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: Coherent information structure in complex computation. Theory in Biosciences **131**(3), 193–203 (2012)
62. Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: Local measures of information storage in complex distributed computation. Information Sciences **208**, 39–54 (2012)
63. Lizier, J.T., Rubinov, M.: Multivariate construction of effective computational networks from observational data. Tech. Rep. Preprint 25/2012, Max Planck Institute for Mathematics in the Sciences (2012)
64. Lungarella, M., Sporns, O.: Mapping Information Flow in Sensorimotor Networks. PLoS Computational Biology **2**(10), e144 (2006)
65. MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge (2003)
66. Mahoney, J.R., Ellison, C.J., James, R.G., Crutchfield, J.P.: How hidden are hidden processes? A primer on crypticity and entropy convergence. Chaos **21**(3), 037,112+ (2011)
67. Manchanda, K., Yadav, A.C., Ramaswamy, R.: Scaling behavior in probabilistic neuronal cellular automata. Physical Review E **87**, 012,704+ (2013)
68. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, USA (1999)
69. Marinazzo, D., Wu, G., Pellicoro, M., Angelini, L., Stramaglia, S.: Information flow in networks and the law of diminishing marginal returns: evidence from modeling and human electroencephalographic recordings. PLoS ONE **7**(9), e45,026 (2012)
70. Mitchell, M.: Computation in Cellular Automata: A Selected Review. In: T. Gramss, S. Bornholdt, M. Gross, M. Mitchell, T. Pellizzari (eds.) Non-Standard Computation, pp. 95–140. VCH Verlagsgesellschaft, Weinheim (1998)
71. Mitchell, M., Crutchfield, J.P., Hraber, P.T.: Evolving Cellular Automata to Perform Computations: Mechanisms and Impediments. Physica D **75**, 361–391 (1994)
72. Nakajima, K., Li, T., Kang, R., Guglielmino, E., Caldwell, D.G., Pfeifer, R.: Local information transfer in soft robotic arm. In: 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1273–1280. IEEE (2012). DOI 10.1109/robio.2012.6491145. URL `http://dx.doi.org/10.1109/robio.2012.6491145`
73. Obst, O., Boedecker, J., Asada, M.: Improving Recurrent Neural Network Performance Using Transfer Entropy Neural Information Processing. Models and Applications. In: K. Wong, B. Mendis, A. Bouzerdoum (eds.) Neural Information Processing. Models and Applications, *Lecture Notes in Computer Science*, vol. 6444, chap. 24, pp. 193–200. Springer Berlin / Heidelberg, Berlin, Heidelberg (2010)
74. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
75. Priesemann, V., Munk, M., Wibral, M.: Subsampling effects in neuronal avalanche distributions recorded in vivo. BMC Neuroscience **10**(1), 40+ (2009)
76. Prokopenko, M., Boschietti, F., Ryan, A.J.: An Information-Theoretic Primer on Complexity, Self-Organization, and Emergence. Complexity **15**(1), 11–28 (2009)
77. Prokopenko, M., Gerasimov, V., Tanev, I.: Evolving Spatiotemporal Coordination in a Modular Robotic System. In: S. Nolfi, G. Baldassarre, R. Calabretta, J. Hallam, D. Marocco, J.A.

Meyer, D. Parisi (eds.) Proceedings of the Ninth International Conference on the Simulation of Adaptive Behavior (SAB'06), Rome, *Lecture Notes in Artificial Intelligence*, vol. 4095, pp. 548–559. Springer Verlag (2006)
78. Prokopenko, M., Lizier, J.T., Obst, O., Wang, X.R.: Relating Fisher information to order parameters. Physical Review E **84**, 041,116+ (2011)
79. Prokopenko, M., Lizier, J.T., Price, D.C.: On thermodynamic interpretation of transfer entropy. Entropy **15**(2), 524–543 (2013)
80. Rubinov, M., Lizier, J., Prokopenko, M., Breakspear, M.: Maximized directed information transfer in critical neuronal networks. BMC Neuroscience **12**(Suppl 1), P18+ (2011)
81. Schreiber, T.: Interdisciplinary application of nonlinear time series methods - the generalized dimensions. Physics Reports **308**, 1–64 (1999)
82. Schreiber, T.: Measuring Information Transfer. Physical Review Letters **85**(2), 461–464 (2000)
83. Shalizi, C.R.: Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata. Ph.D. thesis, University of Wisconsin-Madison (2001)
84. Shalizi, C.R., Haslinger, R., Rouquier, J.B., Klinkner, K.L., Moore, C.: Automatic filters for the detection of coherent structure in spatiotemporal systems. Physical Review E **73**(3), 036,104 (2006)
85. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**, 379–423 & 623–656 (1948)
86. Soon, C.S., Brass, M., Heinze, H.J., Haynes, J.D.: Unconscious determinants of free decisions in the human brain. Nature Neuroscience **11**(5), 543–545 (2008)
87. Staniek, M., Lehnertz, K.: Symbolic transfer entropy. Physical Review Letters **100**(15) (2008). DOI 10.1103/physrevlett.100.158101. URL http://dx.doi.org/10.1103/physrevlett.100.158101
88. Stramaglia, S., Wu, G.R., Pellicoro, M., Marinazzo, D.: Expanding the transfer entropy to identify information subgraphs in complex systems. In: Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3668–3671. IEEE (2012)
89. Ver Steeg, G., Galstyan, A.: Inferring Predictive Links in Social Media Using Content Transfer (2012). URL http://arxiv.org/abs/1208.4475. arXiv:1208.4475
90. Verdes, P.F.: Assessing causality from multivariate time series. Physical Review E **72**(2), 026,222+ (2005)
91. Vicente, R., Wibral, M., Lindner, M., Pipa, G.: Transfer entropy–a model-free measure of effective connectivity for the neurosciences. Journal of Computational Neuroscience **30**(1), 45–67 (2011)
92. Wang, X.R., Miller, J.M., Lizier, J.T., Prokopenko, M., Rossi, L.F.: Quantifying and Tracing Information Cascades in Swarms. PLoS ONE **7**(7), e40,084+ (2012)
93. Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., Lizier, J.T., Vicente, R.: Measuring Information-Transfer delays. PLoS ONE **8**(2), e55,809+ (2013)
94. Wibral, M., Rahm, B., Rieder, M., Lindner, M., Vicente, R., Kaiser, J.: Transfer entropy in magnetoencephalographic data: quantifying information flow in cortical and cerebellar networks. Progress in Biophysics and Molecular Biology **105**(1-2), 80–97 (2011)
95. Williams, P.L., Beer, R.D.: Nonnegative Decomposition of Multivariate Information (2010). URL http://arxiv.org/abs/1004.2515. arXiv:1004.2515
96. Williams, P.L., Beer, R.D.: Generalized Measures of Information Transfer (2011). URL http://arxiv.org/abs/1102.1507. arXiv:1102.1507
97. Wolfram, S.: A New Kind of Science. Wolfram Media, Champaign, IL, USA (2002)
98. Wuensche, A.: Classifying cellular automata automatically: Finding gliders, filtering, and relating space-time patterns, attractor basins, and the Z parameter. Complexity **4**(3), 47–66 (1999)