European Physical Journal B, vol. 73, no. 4, pp. 605-615, 2010. doi: 10.1140/epjb/e2010-00034-5

Differentiating information transfer and causal effect

Joseph T. Lizier[1,2a] and Mikhail Prokopenko[1,3]

[1] CSIRO Information and Communications Technology Centre, Locked Bag 17, North Ryde, NSW 1670, Australia
[2] School of Information Technologies, The University of Sydney, NSW 2006, Australia
[3] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22-26, Leipzig 04103, Germany

Received: December 4, 2009

Abstract. The concepts of information transfer and causal effect have received much recent attention, yet often the two are not appropriately distinguished and certain measures have been suggested to be suitable for both. We discuss two existing measures, transfer entropy and information flow, which can be used separately to quantify information transfer and causal information flow respectively. We apply these measures to cellular automata on a local scale in space and time, in order to explicitly contrast them and emphasize the differences between information transfer and causality. We also describe the manner in which the measures are complementary, including the conditions under which they in fact converge. We show that causal information flow is a primary tool to describe the causal structure of a system, while information transfer can then be used to describe the emergent computation on that causal structure.

PACS. 89.75.Fb Structures and organization in complex systems – 89.75.Kd Patterns – 89.70.Cf Entropy and other measures of information

1 Introduction

Information transfer is currently a popular topic in complex systems science, with recent investigations spanning cellular automata [1], biological signaling networks [2,3], and agent-based systems [4]. In general, information transfer refers to a directional signal or communication of dynamic information from a *source* to a *destination*. However, the body of literature regarding quantification of information transfer appears to subsume two concepts: *predictive* or *computational information transfer*, and *causal effect* or *information flow*. That correlation is not causation is well-understood. Yet while authors increasingly consider the notions of information transfer and information flow and how they fit with our understanding of correlation and causality [5,6,7,8,9,10,11,12], several questions nag. Is information transfer akin to causal effect? If not, what is the distinction between them? When examining the "effect" of one variable on another (e.g. between brain regions), should one seek to measure information transfer or causal effect? Despite the interest in this area, it remains unclear how the notion of information transfer should sit with the concepts of predictive transfer and causal effect.

Predictive transfer refers to the amount of information that a source variable adds to the next state of a destination variable; i.e. "if I know the state of the source, how much does that help to predict the state of the destination?". This *transferred* information can be thought of as adding to the prediction of an observer, or as being transferred into the computation taking place at the destination [13]; as such, we will also refer to this as the computational perspective.

Causal effect refers to the extent to which the source variable has a direct influence or drive on the next state of a destination variable, i.e. "if I change the state of the source, to what extent does that alter the state of the destination?". Information from causal effect can be seen to *flow* through the system, like injecting dye into a river [10]. In an Aristotelian sense, we restrict our interpretation to *efficient* cause here (e.g. see [14]).

Unfortunately, these concepts have become somewhat tangled in discussions of information transfer. Measures for both predictive transfer [9] and causal effect [10] have been inferred to capture information transfer in general, and measures of predictive transfer have been used to infer causality [15,16,11,17] with the two sometimes (problematically) directly equated (e.g. [3,12,5,8,18,19]).

The notion of information transfer remains cloudy while it is used interchangeably to refer to both concepts. Our thesis in this paper is that the concepts of predictive transfer and causal effect are quite distinct: we aim to clarify them and describe the manner in which they should be considered separately. We argue that the concept of predictive transfer (or the computational perspective) is more closely aligned with the popularly understood notion of information transfer, while causal information flow should be considered separately as a useful notion in its own right. Using the perspective of information theory (e.g. see [20]),

[a] jlizier@it.usyd.edu.au

we contend that these concepts are properly quantified by the existing measures known as transfer entropy [9] and information flow [10] respectively, and we use these measures to contrast the concepts.

For this comparison, we examine Cellular Automata (CAs) (e.g. see [21]): discrete dynamical lattice systems involving an array of cells which synchronously update their states as a homogeneous deterministic function of the states of their local neighbors. In particular we focus on Elementary CAs (ECAs), which consist of a one-dimensional array of cells with binary states, with each updated as a function of the previous states of themselves and one neighbor either side (i.e. neighborhood size 3 or range $r = 1$). These previous neighborhood states and the recursive chain of their previous neighborhood states form the *past light-cone* of a cell (i.e. the set of all points capable of having a causal effect on it) [22]. CAs provide a well-known example of complex dynamics, since certain rules (e.g. ECA rules 110 and 54 - see [21] regarding the numbering scheme) exhibit emergent structures which are not discernible from their microscopic update functions but which provide the basis for understanding the macroscopic computations carried out in the CAs [23]. These structures include *particles*, which are coherent structures traveling against a background *domain* region. Regular or periodic particles are known as *gliders*. Particles and gliders are important here because they are popularly understood to embody information transfer in the intrinsic computation in the CA [23].

In particular, we examine the transfer entropy and information flow measures on a *local* scale in space and time in ECAs, in order to provide an explicit comparison between the two. This is the first presentation and examination of the local information flow. We demonstrate that transfer entropy as predictive transfer is more closely aligned with the notion of information transfer (when measured on causal channels), since it alone is associated with emergent coherent information transfer structures, i.e. particles in cellular automata. We also demonstrate that causality stands separately as a useful concept itself, with information flow identifying causal relations in the domain region of the CA and demonstrating the bounds of influence without being confused by correlations. We describe the manner in which these results are generalizable to other systems. Additionally, we describe the conditions and parameter settings under which a variant of the transfer entropy converges with the information flow.

On the basis of these results, we suggest that information flow should be used first wherever possible in order to establish the set of causal information contributors for a given destination variable. Subsequently, the transfer entropy measure may be used to quantify the concept of information transfer from these causal sources (only) to the destination to study emergent computation in the system.

2 Predictive information transfer

2.1 Transfer entropy

The mutual information (e.g. see [20]) measures the average information in a variable Y about another variable X:

$$I(Y;X) = \sum_{y_n, x_n} p(y_n, x_n) \log_2 \frac{p(y_n, x_n)}{p(y_n)p(x_n)}. \quad (1)$$

It was in the past used as a de facto measure of information transfer. Schreiber presented *transfer entropy* [9] in order to address deficiencies in the mutual information, the use of which was criticized in the context of information transfer as a symmetric measure of statically shared information. Transfer entropy is defined as the deviation from independence (in bits) of the state transition of an information destination X from the previous state of an information source Y[1]:

$$T_{Y\to X}(k) = \sum_{w_n} p(w_n) \log_2 \frac{p(x_{n+1}|x_n^{(k)}, y_n)}{p(x_{n+1}|x_n^{(k)})}. \quad (2)$$

Here n is a time index, $x_n^{(k)}$ refers to the k states of X up to and including x_n, and w_n is the state transition tuple $(x_{n+1}, x_n^{(k)}, y_n)$. It can be viewed as a *conditional* mutual information $I(Y; X'|X^{(k)})$, casting it as the average information in the source y_n about the next state of the destination x_{n+1} that was not already contained in the destination's past k states $x_n^{(k)}$. To ensure that no information in the destination's past is mistaken as transfer here, one should take the limit $k \to \infty$ (written $T_{Y\to X}$) though in practice finite-k estimates must be used [1]. This conditioning on the past makes the transfer entropy a *directional*, *dynamic* measure of predictive information, but it remains a measure of observed (conditional) *correlation* rather than direct effect. In fact, the transfer entropy is a nonlinear extension of a concept known as the "Granger causality" [24], the nomenclature for which may have added to the confusion associating information transfer and causal effect.

We note the similar *information current* [25], which measures changes in *spatial* information and does so on a local scale in space and time also. This measure is interpretable as an information contribution between the right and left segments of a system only for those exhibiting deterministic mechanics. Furthermore, in considering spatial information it is only defined for lattice systems, where it either measures information contribution from the right side of the system to the left (or vice-versa) but not from an arbitrary source to an arbitrary destination. As such, it remains out of scope for our comparison of information transfer and causal information flow between a specific source and destination in general multivariate systems.

[1] The transfer entropy can consider transfer from l previous states of the source $y_n^{(l)}$, however here we consider systems where only the previous state of the source is a causal contributor to the destination, so we use $l = 1$.

2.2 Local transfer entropy

The transfer entropy is an average (or *expectation value*) of a *local* transfer entropy [1] at each observation n, i.e. $T_{Y\to X}(k) = \langle t_{Y\to X}(n+1,k)\rangle$ where[2]:

$$t_{Y\to X}(n+1,k) = \log_2 \frac{p(x_{n+1}|x_n^{(k)}, y_n)}{p(x_{n+1}|x_n^{(k)})}. \quad (3)$$

This may also be expressed as a local conditional mutual information: $t_{Y\to X}(n+1,k) = i(y_n; x_{n+1}|x_n^{(k)})$, and written $t_{Y\to X}(n+1)$ in the limit $k\to\infty$.

For lattice systems such as CAs with spatially-ordered elements, the local transfer entropy to variable X_i from X_{i-j} at time $n+1$ is represented as:

$$t(i,j,n+1,k) = \log_2 \frac{p(x_{i,n+1}|x_{i,n}^{(k)}, x_{i-j,n})}{p(x_{i,n+1}|x_{i,n}^{(k)})}. \quad (4)$$

The transfer entropy $t(i,j=1,n+1,k)$ to variable X_i from X_{i-1} at time $n+1$ is illustrated in Fig. 1(a). $t(i,j,n,k)$ is defined for every spatiotemporal destination (i,n), for every information channel or direction j. Sensible values for j correspond to causal information sources, i.e. for CAs, sources within the cell range $|j|\leq r$ (we will see in Section 4.3 that the transfer entropy is interpretable as information transfer for these sources only). We write the average for these lattice systems over i and n as $T(j,k) = \langle t(i,j,n,k)\rangle$.

The transfer entropy may also be conditioned on other possible causal information sources, to eliminate their influence from being attributed to the source in question Y [9]. In general, this means conditioning on all sources Z in X's set of causal information contributors V (except for Y) with joint state $v_{x,n}^y$ at time step n. This gives the local *complete* transfer entropy [1]:

$$t^c_{Y\to X}(n+1,k) = \log_2 \frac{p(x_{n+1}|x_n^{(k)}, y_n, v_{x,n}^y)}{p(x_{n+1}|x_n^{(k)}, v_{x,n}^y)}, \quad (5)$$

$$v_{x,n}^y = \{z_n | \forall Z\in V, Z\neq Y, Z\neq X\}. \quad (6)$$

This too may be expressed as a local conditional mutual information $t^c_{Y\to X}(n+1,k) = i(y_n; x_{n+1}|x_n^{(k)}, v_{x,n}^y)$, and we have $T^c_{Y\to X}(k) = \langle t^c_{Y\to X}(n+1,k)\rangle$.

For CAs this means conditioning on other sources $v_{i,r,n}^j$ within the range r of the destination to obtain [1]:

$$t^c(i,j,n+1,k) = \log_2 \frac{p\left(x_{i,n+1}|x_{i,n}^{(k)}, x_{i-j,n}, v_{i,r,n}^j\right)}{p\left(x_{i,n+1}|x_{i,n}^{(k)}, v_{i,r,n}^j\right)}, \quad (7)$$

$$v_{i,r,n}^j = \{x_{i+q,n} | \forall q : -r\leq q\leq +r, q\neq -j, 0\}. \quad (8)$$

Again, we write $t^c(i,j,n+1)$ in the limit $k\to\infty$ and can express $t^c(i,j,n+1,k) = i(x_{i,n+1}; x_{i-j,n}|x_{i,n}^{(k)}, v_{i,r,n}^j)$.

[2] See Appendix A for consideration of an alternative method of localization of mutual information-based measures

In deterministic systems (e.g. CAs), conditioning on all causal source renders $t^c_{Y\to X}(n+1,k)\geq 0$ because the only possible observed value of x_{n+1} as determined by $\left\{y_n, x_n^{(k)}, v_{x,n}^y\right\}$ has the numerator of the log term in Eq. (5) as $p\left(x_{n+1}|x_n^{(k)}, y_n, v_{x,n}^y\right) = 1$ and a denominator less than or equal to this. Calculations conditioned on no other information contributors (as in Eq. (3)) are labeled as *apparent* transfer entropy, and these values may be positive or negative (see discussion in [1]).

Finally, note that the information (or local entropy) $h(i,n+1) = -\log_2 p(x_{i,n+1})$ required to predict the next state of a destination at time step $n+1$ can be decomposed as a sum of [13]:

- the information gained from the past of the destination (i.e. the mutual information between the past $x_{i,n}^{(k)}$ and next state $x_{i,n+1}$, known as the active information *storage* [13] $a(i,n+1,k) = i(x_{i,n}^{(k)}; x_{i,n+1})$); plus
- the information gained or *transferred* from each causal source considered (in arbitrary order) in the context of that past, incrementally conditioning each contribution on the previously considered sources; plus
- any remaining intrinsic uncertainty in the destination given its past and these sources.

For example, in ECAs we have no intrinsic uncertainty, and the information required to predict the next state of a cell is the sum of information gained from its past, plus the extra information from one source that was not in this past, plus the extra information from the other source that was not in the cell's past or in the first source. The conditional mutual information terms here are equivalent to the apparent transfer entropy from one neighbor plus the complete transfer entropy from the other:

$$h(i,n+1) = a(i,n+1,k) + t(i,j=-1,n+1,k) + t^c(i,j=1,n+1,k). \quad (9)$$

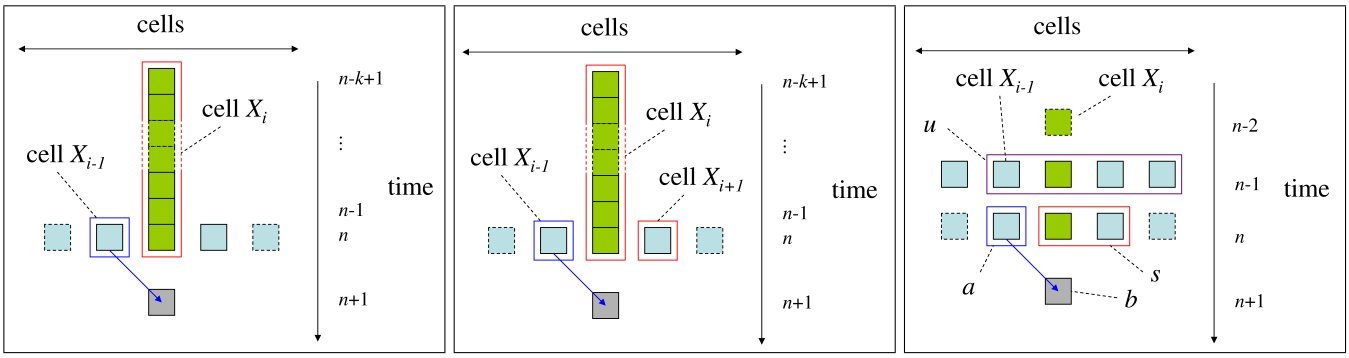
Since the ordering of the sources is arbitrary, we also have:

$$h(i,n+1) = a(i,n+1,k) + t(i,j=1,n+1,k) + t^c(i,j=-1,n+1,k). \quad (10)$$

In this way, the different forms of the transfer entropy as information transfer from causal sources can be seen to characterize important components of the total information at the destination. Importantly, note that no non-causal information sources appear in the sum of information transfer terms contributing to the total information at the destination.

3 Causal effect

It is well-recognized that measurement of causal effect necessitates some type of *perturbation* or *intervention* of the source so as to detect the effect of the intervention on the destination (e.g. see [26,27]). Attempting to infer causality without doing so leaves one measuring correlations of



(a) Apparent transfer entropy (b) Complete transfer entropy (c) Information flow

Fig. 1. Measures of local information transfer and causal information flow across one cell to the right in ECAs. (a) Apparent transfer entropy $t(i, j = 1, n+1, k)$: information contained in the source cell X_{i-1} about the next state of the destination cell X_i at time $n+1$ that was not contained in the destination's past. (b) Complete transfer entropy $t^c(i, j = 1, n+1, k)$: information contained in the source cell X_{i-1} about the next state of the destination cell X_i at time $n+1$ that was not contained in *either* the destination's past *or* the other information contributing cell X_{i+1}. As per Section 4.3, transfer entropy should only be interpreted as information transfer when measured from within the past light-cone of $x_{i,n}$. (c) Information flow $f(i, j = 1, n+1)$: the contribution of a causal effect from source cell X_{i-1} to the next state of the destination cell X_i at time $n+1$, imposing the previous states of the destination cell and the other information contributing cell X_{i+1}; here the source $a = x_{i-1,n}$, the destination $b = x_{i,n+1}$, the imposed contributors are $s = \{x_{i,n}, x_{i+1,n}\}$ and the cells blocking a back-door path (see Section 3.1) relative to (s, a) are $u = \{x_{i-1,n-1}, x_{i,n-1}, x_{i+1,n-1}, x_{i+2,n-1}\}$.

observations, regardless of how directional they may be [10]. Here, we adopt the measure information flow for this purpose, and describe how to apply it on a local scale.

3.1 Information flow

Following Pearl's probabilistic formulation of causal Bayesian networks [26], Ay and Polani [10] consider how to measure causal information flow via *interventional conditional probabilities*. For instance, an interventional conditional probability $p(a|\hat{s})$ considers the distribution of a resulting from *imposing* the value of $\hat{s}$. *Imposing* means intervening in the system to set the value of the imposed variable, and is at the essence of the definition of causal information flow. As an illustration of the difference between interventional and standard conditional probabilities, consider two correlated variables s and a: their correlation alters $p(a|s)$ in general from $p(a)$. If both variables are solely caused by another variable g however, even where they remain correlated we have $p(a|\hat{s}) = p(a)$ because imposing a value $\hat{s}$ has no effect on the value of a.

In a similar fashion to the definition of transfer entropy as the deviation of a destination from *stochastic* independence on the source in the content of the destination's past, Ay and Polani propose the measure *information flow* as the deviation of the destination B from *causal* independence on the source A *imposing* another set of nodes S. Mathematically, this is written as:

$$I_p(A \to B|\hat{S}) = \sum_s p(s) \sum_a p(a|\hat{s}) \sum_b p(b|\hat{a}, \hat{s}) \log_2 \frac{p(b|\hat{a}, \hat{s})}{\sum_{a'} p(a'|\hat{s}) p(b|\hat{a}', \hat{s})}. \quad (11)$$

The value of the measure is dependent on the choice of the set of nodes S. It is possible to obtain a measure of apparent causal information flow $I_p(A \to B)$ from A to B without any S (i.e. $S = \oslash$), yet this can be misleading. For example, it ignores causal information flow arising from interactions of the source with another source variable (e.g. if $b = a$ XOR s and $p(a, s) = 0.25$ for each combination of binary a and s, then $I_p(A \to B) = 0$ despite the clear causal effect of A, while $I_p(A \to B|\hat{S}) = 1$ bit). Also, we may have $I_p(A \to B) > 0$ only because A effects S which in turn effects B; where we are interested in *direct* causal information flow from A to B only $I_p(A \to B|\hat{S})$ validly infers no direct causal effect.

In this paper we are interested in measuring the *direct* causal information flow from A to B, so we must either include all possible other sources in S or at least include enough sources to "block"[3] all non-immediate directed paths from A to B [10]. The minimum to satisfy this is the set of all direct causal sources of B excluding A, including any past states of B that are direct causal sources:

$$s^y_{x,n} = \left\{x_n, v^y_{x,n}\right\}. \quad (12)$$

For computing direct information flow across one cell to the right in ECAs (see Fig. 1(c)) where $a = x_{i-1,n}$ and $b = x_{i,n+1}$, this means S includes the immediate past of the destination cell and the previous state of the cell on its right: $\{x_{i,n}, x_{i+1,n}\}$. Generalized as $I_p(j)$ for information

[3] A set of nodes U *blocks* a path of causal links where there is a node v on the path such that either: i. $v \in U$ and the causal links through v on the path are not both into v, or ii. the causal links through v on the path are both into v and v and all its causal descendants are not in U.

flow across j cells to the right in any 1D CA, we have:

$$s^j_{i,r,n} = \left\{ x_{i,n}, v^j_{i,r,n} \right\}. \quad (13)$$

The major task in computing $I_p(A \to B|\hat{S})$ is the determination of the underlying interventional conditional probabilities in Eq. (11). By definition these may be gleaned by observing the results of intervening in the system, however this is not possible in many cases.

One alternative is to use detailed knowledge of the dynamics, in particular the structure of the causal links and possibly the underlying rules of the causal interactions. This also is often not available in many cases, and indeed is often the very goal for which one turned to such analysis in the first place. Regardless, where such knowledge is available it may allow one to make direct inferences. An important example is where the observed variable is known to be completely determined by the imposing set (e.g. $p(b|\hat{a}, \hat{s})$ in ECAs in Fig. 1(c) can be determined as 0 or 1 from the CA rule table). Indeed, with S selected to compute *direct* information flow, B is determined from A and S (save for any underlying stochasticity), and one can use observational probabilities alone for $p(b|\hat{a}, \hat{s}) = p(b|a, s)$ when all $\{a, s\}$ combinations are observed. Another example is where the observed variable remains unaffected by the imposition (e.g. $p(a|\hat{s})$ becomes $p(a)$ in ECAs in Fig. 1(c)) allowing one to use the observational probabilities alone independently of the imposed variable.

Furthermore, certain other cases exist where one can construct these values from observational probabilities only [10]. For example, the "back-door adjustment" (Section 3.3.1 of [26])[4] is an option where a set of nodes U satisfies the "back-door criteria" relative to (X, Y), i.e. that:

1. no node in U is a causal descendant of X, and
2. U blocks every "back-door path" between X and Y. A back-door path between X and Y is a path of causal links connecting these nodes, where the individual links in the path may point in either direction, so long as the path includes a causal link directly into X. (See footnote 3 for the definition of blocking a path.)

In that case, the interventional conditional probability $p(y|\hat{x})$ is given by:

$$p(y|\hat{x}) = \sum_u p(y|x, u)p(u). \quad (14)$$

The back-door adjustment could be applied to $p(a|\hat{s})$ in ECAs in Fig. 1(c) with the set of nodes satisfying the back-door criteria marked there as u; for $p(b|\hat{a}, \hat{s})$ the set $u_2 = \{u, x_{i-2,n-1}\}$ would be used. In general, note that the back-door adjustment can only be applied if all relevant combinations are observed (i.e. for $\{y, x, u\}$ where $p(y, x, u)$ is strictly positive [10]).

[4] The back-door adjustment is a sub-case of the "adjustment for direct causes" [10] which is numerically simpler when the set of back-door nodes U is known.

3.2 Local information flow

We can define a *local information flow*:

$$f(a \to b|\hat{s}) = \log_2 \frac{p(b|\hat{a}, \hat{s})}{\sum_{a'} p(a'|\hat{s})p(b|\hat{a}', \hat{s})}, \quad (15)$$

in a similar manner to the localization performed for the transfer entropy. The meaning of the local information flow is slightly different however. Certainly, it is an *attribution* of local causal effect of a on b were $\hat{s}$ imposed at the given observation (a, b, s). However, one must be aware that $I_p(A \to B|\hat{S})$ is not the *average* of the local values $f(a \to b|\hat{s})$. Unlike the transfer entropy, the information flow is averaged over a product of *interventional* conditional probabilities ($p(s)p(a|\hat{s})p(b|\hat{a}, \hat{s})$, see Eq. (11)) which in general does not reduce down to the probability of the given observation $p(s, a, b) = p(s)p(a|s)p(b|a, s)$. For instance, it is possible that not all of the tuples $\{a, b, s\}$ will actually be observed, so averaging over observations would ignore the important contribution that any unobserved tuples provide to the determination of information flow. Again, the local information flow is specifically tied not to the given observation at time step n but to the general configuration (a, b, s), and thereby *attributed* to the observation of this configuration at time n.

For lattice systems such as CAs, we use the notation $f(i, j, n+1)$ to denote the local information flow into variable X_i from the source X_{i-j} at time step $n+1$ (i.e. flow across j cells to the right), giving:

$$f(i, j, n+1) = \log_2 \frac{p(x_{i,n+1}|\widehat{x_{i-j,n}}, \widehat{s^j_{i,r,n}})}{d(i, j, n+1)}, \quad (16)$$

$$d(i, j, n+1) = \sum_{x'_{i-j,n}} p(x'_{i-j,n}|\widehat{s^j_{i,r,n}}) \times p(x_{i,n+1}|\widehat{x_{i-j,n}}', \widehat{s^j_{i,r,n}}), \quad (17)$$

with $s^j_{i,r,n}$ defined in Eq. (13).

4 Application to Cellular Automata

Here, we measure the local transfer entropy and local information flow to the raw states of ECA rule 54 in Fig. 2. This rule exhibits a (spatially and temporally) periodic background domain, with gliders traveling across the domain and colliding with one another, forming the basis of an emergent intrinsic computation. We compute the required probabilities from running a 10 000 cell CA for 600 time steps.

Focusing on transfer and flow one step to the right per unit time step, we measure the average transfer values being $T(j = 1, k = 16) = 0.080$ and $T^c(j = 1, k = 16) = 0.193$ bits for apparent and complete transfer entropy respectively, and the information flow at $I_p(j = 1) = 0.523$ bits. Much more insight is provided by examining the *local*

values of each measure however, and we describe four cases within these results to highlight the differences in the measures and indeed in the concepts of information transfer and causal effect in general. These differences have been observed for transfer and flow one step to the left per unit time step (i.e. $j = -1$) also, and in other CAs with emergent structure (e.g. rules 110 and 18), and we comment on the generality of these results to other systems.

For the information flow, $p(b|\hat{a}, \hat{s})$ is measured using observations only as per the back-door adjustment described in Section 3.1 (unless otherwise stated), to minimize reliance on knowledge of the underlying dynamics.

4.1 Coupled periodic processes may be highly causal

As an extension of the example of coupled Markov chains in [10] to more complex dynamics, we first look at the background domain region of the CA where each cell executes a periodic sequence of states. The four time step period of the (longest) sequences is longer than any one binary-state cell could produce alone – the cells rely on interaction with their neighbors to produce these long sequences. We see that the local transfer entropies $t(i, j = 1, n, k = 16)$ and $t^c(i, j = 1, n, k = 16)$ vanish here in Fig. 2(d) and Fig. 2(e) [1], while the local information flow $f(i, j = 1, n)$ in Fig. 2(b) measures a periodic pattern of causal effect at similar levels to those in the glider/blinker regions.

Both results are correct, but from different perspectives. From a computational perspective, the cells in the domain region are executing *information storage* processes – their futures are (almost) completely predictable from their periodic pasts alone [1] (i.e. $p(x_{i,n+1}|x^{(k)}_{i,n}, x_{i-j,n}) \to p(x_{i,n+1}|x^{(k)}_{i,n})$) so there is vanishing information transfer here (and $t(i, j = 1, n, k = 16) \to 0$). This is clearly the case for any periodic process at the destination. Note that to achieve the long periods here, some information is stored in neighbors and retrieved after a few time steps [13] (achieving a stigmergic information storage, similar to [28]). Indeed, the long periodic sequences in the background domain (longer than any one cell could produce alone) are *necessarily* underpinned by the coupled causal effect between the neighbors. From another perspective, much of the background domain is highly causal simply because had one imposed values on the sources there the destinations would have changed; hence we find the strong patterns of information flow here.

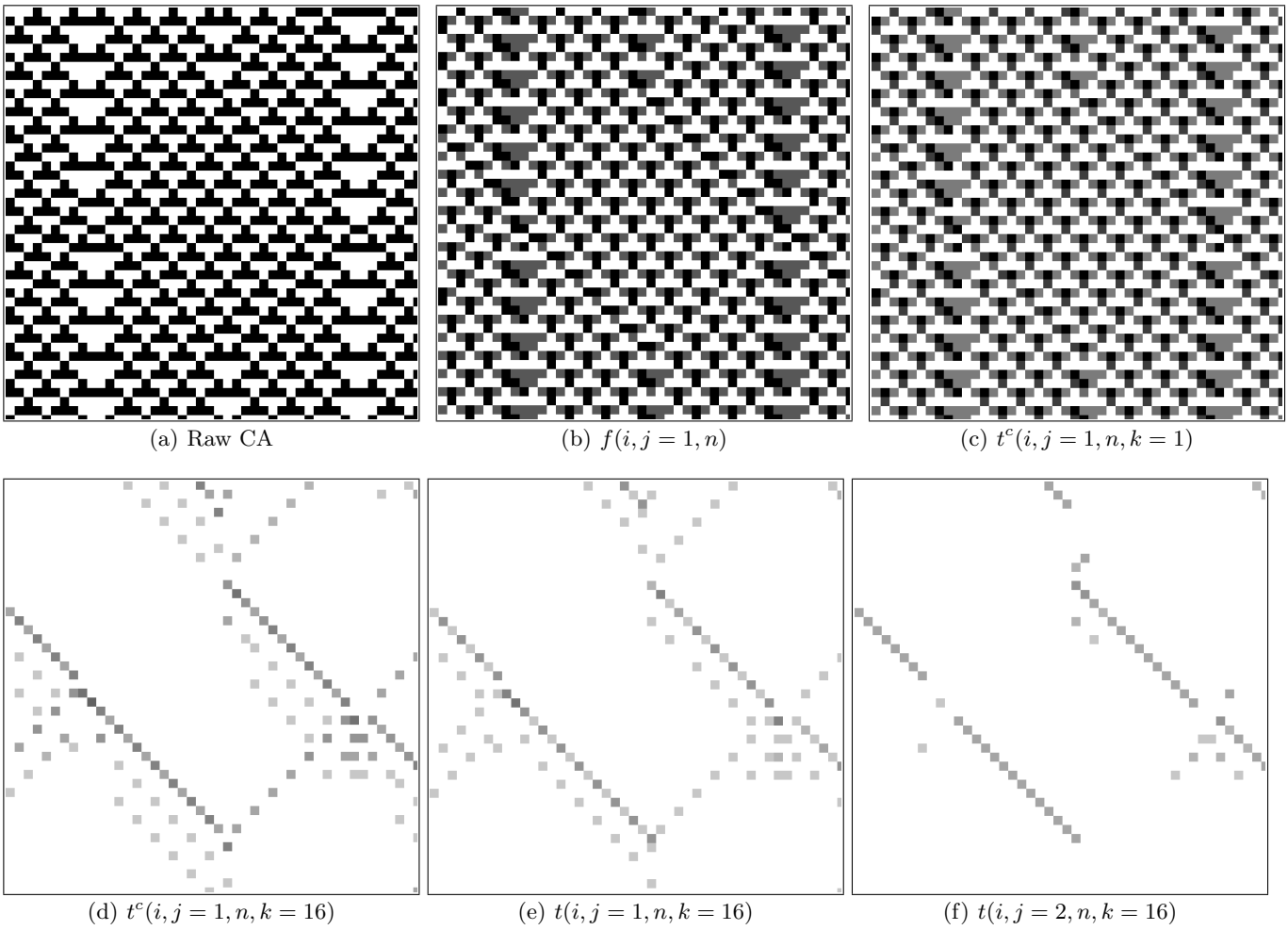
The key general result here is that the measure *transfer entropy does not detect all causal effects that information flow does.* This is because the concept of information transfer is focused on computation and is not intended to capture causal effect where that causal effect underpins information storage instead.

4.2 Gliders distinguished as emergent information transfer

We then examine the measurements at the gliders, the emergent structures which propagate against the background domain. Here we see that the local transfer entropies $t(i, j = 1, n, k = 16)$ and $t^c(i, j = 1, n, k = 16)$ measure strong predictive information in the direction of glider motion in Fig. 2(d) and Fig. 2(e) [1], while the local information flow $f(i, j = 1, n)$ measurement in Fig. 2(b) varies little between the gliders and the background domain.

Again, *both results are correct from different perspectives.* The cell states in the glider region provide much stronger predictive information about the next states in the direction of glider motion than do the previous states of the destination cells (i.e. $p(x_{i,n+1}|x^{(k)}_{i,n}, x_{i-j,n}) > p(x_{i,n+1}|x^{(k)}_{i,n})$). This is why gliders have long been said to transfer information about the dynamics in one part of the CA to another (as quantified by the local transfer entropy [1]). Indeed in Eq. (9) we see these as information transfer terms combining with information storage in computing the next state of the cell. For these reasons, we say that predictive transfer is the concept that more closely aligned with the popularly understood concept of information transfer. From a causal perspective, the same CA rules or templates $\{a, s\}$ executed in the glider are also executed elsewhere in the domain of the CA – while imposing the source value does indeed have a causal effect on the destination in the gliders, the positive directional information flow here is no greater than levels observed in the domain. The measure certainly captures the causal mechanism in the gliders, but its localization does not distinguish that from the flow in the domain.

The key general result then is that *information flow does not distinguish emergent computational structure (i.e. gliders) that transfer entropy does.* It is possible that a macroscopic formulation of the information flow might distinguish gliders as highly causal macroscopic structures, but certainly (when applied to the same source and destination pair as transfer entropy) as a directional measure of direct local causal effect it does not distinguish these emergent structures. In this form, the causal perspective focuses on the details or micro-level of the dynamics, whereas the predictive or computational perspective takes a macroscopic view of emergent structures. It is the examination in the context of the past k states that affords this macroscopic view to the transfer entropy, and emergent structure can only be detected on this scale. Since gliders are dislocations in background patterns [29] which can only be caused by neighboring cells, the source of the glider will add information about the destination in the context of this pattern in the past k states (i.e. $p(x_{i,n+1}|x^{(k)}_{i,n}, x_{i-j,n}) > p(x_{i,n+1}|x^{(k)}_{i,n})$) and we have strong information transfer. On the other hand, information flow intrinsically cannot consider the context of the past, since imposing on $x_{i-j,n}$ and $s^j_{i,r,n}$ blocks out the influence of those past k states.



(a) Raw CA (b) $f(i, j = 1, n)$ (c) $t^c(i, j = 1, n, k = 1)$

(d) $t^c(i, j = 1, n, k = 16)$ (e) $t(i, j = 1, n, k = 16)$ (f) $t(i, j = 2, n, k = 16)$

Fig. 2. Local transfer entropy and information flow for raw states of rule 54 in (a) (45 time steps displayed for 45 cells, time increases down the page): (b) *Local information flow* across one cell to the right, (all figures gray-scale with 16 levels) with max. 1.07 bits (black); *Local complete transfer entropy* across one cell to the right: (c) with past history length $k = 1$ (max. 1.17 bits (black)), and (d) past history length $k = 16$ (max. 9.22 bits (black)); *Local apparent transfer entropy*, positive values only: (e) across one cell to the right (max. 7.93 bits (black)), and (f) across two cells to the right (max. 6.00 bits (black)).

Finally, we note that the highlighting of the gliders by the local transfer entropy is a form of filtering for emergent structure in CAs [1]. It is not the only method of such filtering (see also: finite-state transducers which recognize the regular spatial language of the CA [29,30]; local statistical complexity and local sensitivity [22]; displays of neighborhood rule templates with the most frequently occurring rules filtered out [31]; and local information, i.e. local spatial entropy rate [32]). However, it is the only method to do so with an explicit measure of information transfer and to provide direct quantitative evidence that gliders are information transfer agents.

4.3 Information transfer to be measured from causal sources only

Fig. 2(f) measures the local apparent transfer entropy $t(i, j = 2, n, k = 16)$ for two steps to the right per unit time step. This profile is fairly similar to that produced for one step to the right per unit time step (Fig. 2(e)). However, this measurement suggests a *superluminal* transfer, i.e. transfer from outside of the past light-cone of $x_{i,n}$. The result is not intuitive as we expect zero information transfer from sources that are not direct causal information contributors. This is because only causal sources are present in Eq. (9) in contributing or *transferring* information to the next state of the destination. What we see in this profile merely reflects a correlation between the purported source and an actual causal source one cell away from the destination: the transfer entropy will produce a non-zero result from non-causal sources whenever such correlations exist. This does not mean that the transfer entropy measure is wrong, merely that it has not been correctly interpreted here. The key general result is that *in order to be genuinely interpreted as information transfer, the transfer entropy should only be applied to causal information sources for the given destination.* Beyond these

sources, it only measures correlations that do not directly contribute or transfer information into the computation that determines the next state of the destination.

To check the correctness of the information flow measure, we apply it here assuming the CA is of neighborhood-5 (i.e. two causal contributors on either side of the destination with $r = 2$). As expected, the local information flow profile computes no causal effect across two cells to the right per unit time step (not shown). Importantly however, note that the information flow could not be measured using observational data alone for either $j = 1$ or $j = 2$ in neighborhood-5 (since the CA does not produce all of the required $\{s, a\}$ combinations for computing $p(b|\hat{a}, \hat{s})$); specific knowledge about the dynamics was required for the calculation.

Furthermore, measuring the complete transfer entropy $t^c(i, j = 2, n, k = 16)$ in this neighborhood results in a zero predictive information profile (not shown). This is because all the information $h(i, n+1)$ required to predict the next state of the destination is contained within the interior $r = 1$ neighborhood for this deterministic system. This information is represented in the right-hand side of Eq. (9), and is a subset of the conditioned variables in the expansion of $t^c(i, j = 2, n, k = 16)$ as $i(x_{i,n+1}; x_{i-2,n}|x^{(k)}_{i,n}, v^{j=2}_{i,r=2,n})$. This measurement aligns well with the zero result for information flow. Significantly, only the complete transfer entropy is able to make its inference using the available observational data alone, though both measures require the correct neighborhood of other causal contributors to be a subset of those conditioned on or imposed here.

4.4 Complete transfer entropy as an inferrer for information flow

The parallels between the complete transfer entropy and the information flow go beyond similar inference of a lack of influence. Consider the profile of $t^c(i, j = 1, n, k = 1)$ in Fig. 2(c) – note how similar it is to the profile of the local information flow in Fig. 2(b). Note also that the average value $T^c(j = 1, k = 1) = 0.521$ bits is almost identical to the information flow $I_p(j = 1) = 0.523$ bits.

Convergence of the complete transfer entropy and direct information flow occurs with the combination of *one* parameter setting and *two* conditions which are approximated in this example:

1. the parameter k for the complete transfer entropy was set to include only the past states of the destination that are causal information contributors to its next state;
2. the condition that all $\{a, s\}$ combinations are observed (this condition is relevant for averages but not local values);
3. the condition that $p(a|\hat{s}) \equiv p(a|s)$ (which for example is met where a is both causally and conditionally independent of s).

We describe why these conditions lead to convergence in the following paragraphs.

With history length $k = 1$ here the numerators of the local measures Eq. (5) and Eq. (15) in fact become equal. This is enabled because with the history length k set to include only the past states of the destination that are causal information contributors to its next state[5] – *no more, no less* – the complete transfer entropy *conditions* on the same variables that the direct information flow *imposes* upon. That is, as shown in Eq. (12) $s^y_{x,n}$ for Eq. (15) refers to the same variables as $\left\{x^{(k)}_n, v^y_{x,n}\right\}$ in Eq. (5) with k set in this manner. Building on this enabling then, since we are measuring *direct* information flow we have $p(b|\hat{a}, \hat{s}) = p(b|a, s)$ when the $\{a, s\}$ combination is observed (as stated in Section 3.1). This parameter setting then ensures the numerators of the *local* measures Eq. (5) and Eq. (15) are the same.

Note that for the combinations of $\{a, s\}$ which are not observed $p(b|a, s)$ is technically undefined; this is not relevant for local values of either measure (since the given $\{a, s\}$ must have been observed), or the average complete transfer entropy, but for the information flow these terms revert to $p(b|\hat{a}, \hat{s})$ and contribute additionally to I_p. For convergence of the *averages* $T^c_{Y\to X}(k)$ and $I_p(A \to B|\hat{S})$ only, it is thus required that all $\{a, s\}$ combinations are observed. The condition is met in this example.

Consider now that if the condition $p(a|\hat{s}) \equiv p(a|s)$ is also met, then the denominator of Eq. (15) becomes $p(b|s)$. With s referring to the same variables as $\left\{x^{(k)}_n, v^y_{x,n}\right\}$, the denominator of Eq. (15) then matches Eq. (5), and in conjunction with the above conditions we have equality between the two local values and their averages. Importantly, this condition does not require all values of s to be observed for convergence of the averages, since $p(s)$ in Eq. (11) eliminates the contribution of any unobserved values of s.

This final condition is approximated but not quite exactly met in the CA example. As described in Section 3.1, we have $p(a|\hat{s}) \equiv p(a)$ for the information flow here. We note that this final condition would still be met if in fact $p(a) = p(a|s)$ (i.e. $p(y_n|x^{(k)}_n, v^y_{x,n}) = p(y_n)$ in the notation for Eq. (5)). That is, there is a class of systems which satisfy this condition because the source is *both* causally and conditionally independent of the other causal contributors to the destination. The CA example approximates the sub-condition $p(a|s) = p(a)$. In Fig. 1 we see that while both $a = x_{i-1,n}$ and $s = \{x_{i,n}, x_{i+1,n}\}$ have two common sources ($\{x_{i-1,n-1}, x_{i-1,n}\}$), a has one extra and s has two extra sources that are not shared. It is these unshared sources that cannibalize the correlation between s and a. The small correlation here is confirmed by the Kullback-Leibler divergence (see [20]) of $p(a|s)$ from $p(a)$ (i.e. the mutual information between a and s) which is very low (0.01 bits) for $j = \{1, -1\}$ for rule 54 here. The divergence is still low, but larger for other ECA rules with emergent

[5] That is, with $k = 1$ in the CAs here, though for example in [11] where the elements in Henon maps are causally effected by their previous two states, $k = 2$ would be appropriate rather than the use of $k = 1$ there.

structure (i.e. 0.03 bits for rule 110 and 0.13 bits for rule 18). Nonetheless, the non-zero divergence confirms that the condition is not precisely met. Finally we note that where the previous conditions (including $p(a|\hat{s}) \equiv p(a)$) were met, the difference between the local values due to $p(a|s) \neq p(a)$ may be written as:

$$\log_2 \frac{\sum_{a'} p(a')p(b|a',s)}{p(b|s)}, \tag{18}$$

or for the CA as:

$$\log_2 \frac{\sum_{x'_{i-j,n}} p(x'_{i-j,n})p(x_{i,n+1}|x'_{i-j,n}, s^j_{i,r,n})}{p(x_{i,n+1}|s^j_{i,r,n})}. \tag{19}$$

Interestingly this difference is independent of the source value, $a = x_{i-j,n}$.

Where one cannot intervene in the system, and does not have the required observations to use a method such as the back-door adjustment, the local complete transfer entropy could provide a useful inference for the local information flow profile. Within one's control is to set the history length k to include only the past states of the destination that are causal information contributors to its next state. The history length parameter k therefore has an important role in moving the (complete) transfer entropy between measuring information transfer (at large k) and approximating causal effect (at minimal k). Outside of one's control is whether the other conditions are met; errors begin to be introduced where they are not. We note that there is a wide class of systems where the source a is causally independent of the other causal contributors to the destination s (i.e. $p(a|\hat{s}) \equiv p(a)$), and though error-prone a subsequent assumption of conditional independence (i.e. $p(a|s) = p(a)$) is a maximum entropy assumption.

Importantly, the complete transfer entropy must condition on the correct neighborhood of causal sources. This knowledge is missing in the important application where one is *inferring* causal structure in a multi-variate time series. It is possible that the transfer entropy itself could be used to iteratively *build* an inference of the causal contributors for a given destination by *incrementally* conditioning on previously inferred sources (reminiscent of Eq. (9)). This would be done by incrementally identifying the next source which provides the most statistically significant transfer entropy conditioned on the previously identified sources, until all (deterministic) information in the destination is accounted for. Such a method combines the multi-variate source selection of [3] with the complete transfer entropy and the statistical significance tests of [17]. Testing this method is left for future work.

Finally, we note that while the complete transfer entropy can at least function in the absence of observations spanning all possible combinations of the variables, if crucial combinations are not observed it can give quite incorrect inferences here. For example, consider the classical causal example of a short circuit which causes a fire in the presence of certain conditions (e.g. with inflammable material), while the fire can also be started in other ways (e.g. overturning a lighted oil stove) [33]. If one never observes the short circuit in the right conditions, without the other fire triggers, the transfer entropy is in fact unable to infer a causal link from the short circuit to the fire.

5 Discussion and Conclusion

The concepts of information transfer and causal effect have often been confused. In this paper, we have demonstrated the complementary nature of these concepts while emphasizing the distinctions between them. On an information-theoretical basis, information flow quantifies causal effect using an interventionist perspective, while transfer entropy quantifies information transfer by measuring a (conditional) correlation on a causal channel. We have explored the subtle yet distinct differences between these concepts using a local scale within cellular automata.

Causal effect is a fundamental micro-level property of a system. Information flow should be used as a primary tool (where possible) to establish the presence of and quantify causal relationships. There are situations where this is not possible (e.g. where one has no ability to intervene in the system, no knowledge of the underlying dynamics, and cannot apply a method such as the back-door adjustment to observational data). Then, under certain parameter settings (i.e. with history length k set to include only the causal contributors from the destination's past) and conditions the complete transfer entropy converges with the information flow, and may still provide a reasonable inference where these conditions are approximated. The apparent transfer entropy is not applicable here since it cannot discern correlation from causal effect, and neither apparent nor complete transfer entropy with large k is suitable since these measure predictive information transfer rather than direct causal effect. Note that for both the information flow or complete transfer entropy, it is crucial that they be applied imposing or conditioning the correct set of other causal variables – the task of building knowledge of this correct set is left for investigation in future work.

Information transfer can then be analyzed in order to gain insight into the emergent computation being carried out by the system, e.g. via gliders in CAs. Importantly, the transfer entropy should only be measured for causal information contributors to the destination, otherwise its result cannot be interpreted as information transfer. To do so, both the apparent and complete transfer entropy should be used, with history length k set as large as possible. These are complementary measures which allow one to assess the composition of information storage, transfer and interactions in a system [13]. Information flow is not suitable for the analysis of emergent computation, since in representing causal effect it takes too microscopic a viewpoint, and provides no method for describing the composition of information in the computation.

The authors thank Daniel Polani and Nihat Ay for helpful discussions regarding the nature of the information flow measure

and in particular how to estimate it from observational data. JL thanks John Mahoney for discussions regarding measuring transfer entropy from non-causal sources, and the Australian Research Council Complex Open Systems Research Network (COSNet) for a travel grant that partially supported this work.

A Consideration of alternative method of localization

An alternative method of localizing mutual information-based measures was proposed in [34]. The authors consider *partial* localizations, computing how much information $I(y_n; X)$ a specific value y_n gives about what value X *might* take. It is required that a partial localization $I(y_n; X)$ averages over y_n to the average mutual information $I(Y; X)$:

$$I(Y; X) = \sum_{y_n} p(y_n) I(y_n; X). \tag{20}$$

As well as the conventional expression that satisfies this requirement:

$$I_1(y_n; X) = \sum_{x_n} p(x_n|y_n) \log_2 \frac{p(x_n|y_n)}{p(x_n)}, \tag{21}$$

the authors present an alternative partial local mutual information as the reduction in uncertainty of X on knowing y_n:

$$I_2(y_n; X) = H_X - H_{X|y_n}, \tag{22}$$

giving:

$$\begin{aligned} I_2(y_n; X) = & -\sum_{x_n} p(x_n) \log_2 p(x_n) + \\ & \sum_{x_n} p(x_n|y_n) \log_2 p(x_n|y_n). \end{aligned} \tag{23}$$

While both I_1 and I_2 satisfy the constraint Eq. (20), they do give different values for $I(y_n; X)$. Importantly, I_1 is *non-negative*, but I_2 is unique in satisfying the key property of *additivity* of information from multiple sources:

$$I(\{y_n, z_n\}; X) = I(y_n; X) + I(z_n; X \mid y_n). \tag{24}$$

In this paper we consider *full* localizations, computing how much information $i(y_n; x_n)$ a value y_n gives about the specific value x_n that X *actually* takes at time step n. Similar to requirement Eq. (20), the full localizations $i(y_n; x_n)$ are required to satisfy:

$$I(Y; X) = \sum_{y_n} p(y_n) \sum_{x_n} p(x_n|y_n) i(y_n; x_n). \tag{25}$$

The approach to these local values used in the main body of our text:

$$i_1(y_n; x_n) = \log_2 \frac{p(x_n|y_n)}{p(x_n)}, \tag{26}$$

is analogous to $I_1(y_n; X)$ because it also satisfies:

$$I(y_n; X) = \sum_{x_n} p(x_n|y_n) i(y_n; x_n), \tag{27}$$

for $I_1(y_n; X)$. Interestingly, for $i_1(y_n; x_n)$ we also have:

$$i_1(y_n; x_n) = h(x_n) - h(x_n|y_n), \tag{28}$$

in analogy to $I_2(y_n; X)$ in Eq. (22), which leads i_1 to satisfy the crucial property of *additivity* [34]:

$$i(\{y_n, z_n\}; x_n) = i(y_n; x_n) + i(z_n; x_n \mid y_n), \tag{29}$$

unlike $I_1(y_n; X)$ (with Eq. (24)).

It is worth considering whether the approach of [34] in proposing $I_2(y_n; X)$ may be extended to propose a valid $i_2(y_n; x_n)$ which satisfies Eq. (25) by satisfying Eq. (27) for $I_2(y_n; X)$. Certainly an extension of Eq. (23) provides:

$$i_2(y_n; x_n) = -\frac{p(x_n)p(y_n)}{p(x_n, y_n)} \log_2 p(x_n) + \log_2 p(x_n|y_n), \tag{30}$$

for this purpose. However, this expression does not satisfy the additivity property of Eq. (29).

Importantly also, expressions for $i(y_n; x_n)$ have an additional requirement for correctness: they **must** be *symmetric* in x_n and y_n in analogy to the averaged value $I(X; Y)$ because the information contained in y_n about the specific value x_n is the same as the information contained in x_n about the specific value of y_n. This is not applicable to partial localizations $I(y_n; X)$ because they are asymmetrically defined in considering the known value of one variable and the unknown value of the other. The extension of $I_2(y_n; X)$ to $i_2(y_n; x_n)$ fails this symmetry requirement in general (easily verified with sample values, e.g. $p(x_n) = 0.1$, $p(y_n) = 0.18$, $p(x_n|y_n) = 0.5$, $p(y_n|x_n) = 0.9$, $p(x_n, y_n) = 0.09$), and so is not a correct form to locally quantify the mutual information.

As such, we are left with $i_1(y_n; x_n)$ for *full* localizations $i(y_n; x_n)$ since *it satisfies both additivity and symmetry.*

When selecting a measure for *partial* localizations, one should carefully consider which properties are required. Selecting $I_2(y_n; X)$ preserves additivity, while $I_1(y_n; X)$ preserves positivity and averaging over the correct full localization $i_1(y_n; x_n)$.

References

1. J.T. Lizier, M. Prokopenko, A.Y. Zomaya, Phys. Rev. E **77**(2), 026110 (2008)
2. J. Pahle, A.K. Green, C.J. Dixon, U. Kummer, BMC Bioinformatics **9**, 139 (2008)
3. T.Q. Tung, T. Ryu, K.H. Lee, D. Lee, *Inferring Gene Regulatory Networks from Microarray Time Series Data Using Transfer Entropy*, in *Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS '07), Maribor, Slovenia*, edited by P. Kokol, V. Podgorelec, D. Mičetič-Turk, M. Zorman, M. Verlič (IEEE, Los Alamitos, 2007), pp. 383–388

4. M. Lungarella, O. Sporns, PLoS Comput. Biol. **2**(10), e144 (2006)
5. X.S. Liang, Phys. Rev. E **78**(3), 031113 (2008)
6. N. Lüdtke, S. Panzeri, M. Brown, D.S. Broomhead, J. Knowles, M.A. Montemurro, D.B. Kell, J. R. Soc. Interface **5**(19), 223 (2008)
7. G. Auletta, G.F.R. Ellis, L. Jaeger, J. R. Soc. Interface **5**(27), 1159 (2008)
8. K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, J. Bhattacharya, Physics Reports **441**(1), 1 (2007)
9. T. Schreiber, Phys. Rev. Lett. **85**(2), 461 (2000)
10. N. Ay, D. Polani, Adv. Complex Syst. **11**(1), 17 (2008)
11. M. Lungarella, K. Ishiguro, Y. Kuniyoshi, N. Otsu, Int. J. Bifurcation Chaos **17**(3), 903 (2007)
12. K. Ishiguro, N. Otsu, M. Lungarella, Y. Kuniyoshi, Phys. Rev. E **77**(2), 026216 (2008)
13. J.T. Lizier, M. Prokopenko, A.Y. Zomaya, *A framework for the local information dynamics of distributed computation in complex systems* (2008), arXiv:0811.2690, `http://arxiv.org/abs/0811.2690`
14. H.B. Veatch, *Aristotle: A contemporary appreciation* (Indiana University Press, Bloomington, 1974)
15. H. Sumioka, Y. Yoshikawa, M. Asada, *Causality Detected by Transfer Entropy Leads Acquisition of Joint Attention*, in *Proceedings of the 6th IEEE International Conference on Development and Learning (ICDL 2007), London* (IEEE, 2007), pp. 264–269
16. M. Vejmelka, M. Palus, Phys. Rev. E **77**(2), 026214 (2008)
17. P.F. Verdes, Phys. Rev. E **72**(2), 026222 (2005)
18. G. Van Dijck, J. Van Vaerenbergh, M.M. Van Hulle, *Information Theoretic Derivations for Causality Detection: Application to Human Gait*, in *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2007), Porto, Portugal*, edited by J.M.d. Sá, L.A. Alexandre, W. Duch, D. Mandic (Springer-Verlag, Berlin/Heidelberg, 2007), Vol. 4669 of *Lecture Notes in Computer Science*, pp. 159–168
19. Y.C. Hung, C.K. Hu, Phys. Rev. Lett. **101**(24), 244102 (2008)
20. D.J. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003)
21. S. Wolfram, *A New Kind of Science* (Wolfram Media, Champaign, IL, USA, 2002)
22. C.R. Shalizi, R. Haslinger, J.B. Rouquier, K.L. Klinkner, C. Moore, Phys. Rev. E **73**(3), 036104 (2006)
23. M. Mitchell, in *Non-Standard Computation*, edited by T. Gramss, S. Bornholdt, M. Gross, M. Mitchell, T. Pellizzari (VCH Verlagsgesellschaft, Weinheim, 1998), pp. 95–140
24. C.W.J. Granger, Econometrica **37**, 424 (1969)
25. T. Helvik, K. Lindgren, M.G. Nordahl, Comm. Math. Phys. **272**(1), 53 (2007)
26. J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, 2000)
27. L.R. Hope, K.B. Korb, Tech. Rep. 2005/176, Clayton School of Information Technology, Monash University (2005)
28. A.S. Klyubin, D. Polani, C.L. Nehaniv, *Tracking Information Flow through the Environment: Simple Cases of Stigmergy*, in *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALife IX), Boston, USA*, edited by J. Pollack, M. Bedau, P. Husbands, T. Ikegami, R.A. Watson (MIT Press, Cambridge, MA, USA, 2004), pp. 563–568
29. J.E. Hanson, J.P. Crutchfield, J. Stat. Phys. **66**, 1415 (1992)
30. J.E. Hanson, J.P. Crutchfield, Physica D **103**(1-4), 169 (1997)
31. A. Wuensche, Complexity **4**(3), 47 (1999)
32. T. Helvik, K. Lindgren, M.G. Nordahl, *Local information in one-dimensional cellular automata*, in *Proceedings of the International Conference on Cellular Automata for Research and Industry, Amsterdam*, edited by P.M. Sloot, B. Chopard, A.G. Hoekstra (Springer, Berlin/Heidelberg, 2004), Vol. 3305 of *Lecture Notes in Computer Science*, pp. 121–130
33. J.L. Mackie, in *Causation*, edited by E. Sosa, M. Tooley (Oxford University Press, New York, USA, 1993)
34. M.R. DeWeese, M. Meister, Network: Computation in Neural Systems **10**, 325 (1999)