Anatomy of a Bit:
Information in a Time Series Observation

Ryan G. James,[1, 2, *] Christopher J. Ellison,[1, 2, †] and James P. Crutchfield[1, 2, 3, ‡]

[1] *Complexity Sciences Center*
[2] *Physics Department*
University of California at Davis,
One Shields Avenue, Davis, CA 95616
[3] *Santa Fe Institute*
1399 Hyde Park Road, Santa Fe, NM 87501
(Dated: August 18, 2011)

Appealing to several multivariate information measures—some familiar, some new here—we analyze the information embedded in discrete-valued stochastic time series. We dissect the uncertainty of a single observation to demonstrate how the measures' asymptotic behavior sheds structural and semantic light on the generating process's internal information dynamics. The measures scale with the length of time window, which captures both intensive (rates of growth) and subextensive components. We provide interpretations for the components, developing explicit relationships between them. We also identify the informational component shared between the past and the future that is not contained in a single observation. The existence of this component directly motivates the notion of a process's effective (internal) states and indicates why one must build models.

Keywords: entropy, total correlation, multivariate mutual information, binding information, entropy rate, predictive information rate

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey 02.50.Ga

A single measurement, when considered in the context of the past and the future, contains a wealth of information, including distinct kinds of information. Can the present measurement be predicted from the past? From the future? Or, only from them together? Or not at all? How much of the measurement value is due to randomness? Does that randomness have consequences for the future or it is simply lost? We answer all of these questions and more, giving a complete dissection of a measured bit of information.

I.   INTRODUCTION

In a time series of observations, what can we learn from just a single observation? If the series is a sequence of coin flips, a single observation tells us nothing of the past nor of the future. It gives a single bit of information about the present—one bit out of the infinite amount the time series contains. However, if the time series is periodic—say, alternating 0s and 1s—then with a single measurement in hand, the entire observation series need not be stored; it can be substantially compressed. In fact, a single observation tells us the oscillation's phase. And, with this single bit of information, we have learned *everything*—the full bit that the time series contains. Most systems fall somewhere between these two extremes. Here, we develop an analysis of the information contained in a single measurement that applies across this spectrum.

Starting from the most basic considerations, we deconstruct what a measurement is, using this to directly step through and preview the main results. With that framing laid out, we reset, introducing and reviewing the relevant tools available from multivariate information theory including several that have been recently proposed. At that point, we give a synthesis employing information measures and the graphical equivalent of the information diagram. The result is a systematic delineation of the kinds of information that the distribution of single measurements can contain and their required contexts of interpretation. We conclude by indicating what is missing in previous answers to the measurement question above, identifying what they do and do not contribute, and why alternative state-centric analyses are ultimately more comprehensive.

II.   A MEASUREMENT: A SYNOPSIS

For our purposes an instrument is simply an interface between an observer and the system to which it attends. All the observer sees is the instrument's output—here, we

* rgjames@ucdavis.edu
† cellison@cse.ucdavis.edu
‡ chaos@cse.ucdavis.edu

take this to be one of k discrete values. And, from a series of these outputs, the observer's goal is to infer and to understand as much about the system as possible—how predictable it is, what are the active degrees of freedom, what resources are implicated in generating its behavior, and the like.

The first step in reaching the goal is that the observer must store at least one measurement. How many decimal digits must its storage device have? To specify which one of k instrument outputs occurred the device must use $\log_{10} k$ decimal digits. If the device stores binary values, then it must provide $\log_2 k$ bits of storage. This is the maximum for a one-time measurement. If we perform a series of n measurements, then the observer's storage device must have a capacity of $n \log_2 k$ bits.

Imagine, however, that over this series of measurements it happens that output 1 occurs n_1 times, 2 occurs n_2 times, and so on, with k occurring n_k times. It turns out that the storage device can have much less capacity; using less, sometimes substantially less, than $n \log_2 k$ bits.

To see this, recall that the number M of possible sequences of n measurements with $n_1, n_2, \ldots, n_k$ counts is given by the multinomial coefficient:

$$
\begin{aligned}
M &= \binom{n}{n_1 \ n_2 \ \cdots \ n_k} \\
&= \frac{n!}{n_1! \cdots n_k!} \ .
\end{aligned}
$$

So, to specify which sequence occurred we need no more than:

$$
k \log_2 n + \log_2 M + \log_2 n + \cdots
$$

The first term is the maximum number of bits to store the count n_i of each of the k output values. The second term is the number of bits needed to specify the particular observed sequence within the class of sequences that have counts $n_1, n_2, \ldots, n_k$. The third term is the number b of bits to specify the number of bits in n itself. Finally, the ellipsis indicates that we have to specify the number of bits to specify b ($\log_2 \log_2 n$) and so on, until there is less than one bit.

We can make sense of this and so develop a helpful comparison to the original storage estimate of $n \log_2 k$ bits, if we apply Stirling's approximation: $n! \approx \sqrt{2\pi n}\,(n/e)^n$. For a sufficiently long measurement series, a little algebra gives:

$$
\begin{aligned}
\log_2 M &\approx -n \sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n} \\
&= nH[n_1/n, n_2/n, \ldots, n_k/n] \ .
\end{aligned}
$$

bits for n observations. Here, the function $H[P]$ is Shannon's *entropy* of the distribution $P = (n_1/n, n_2/n, \ldots, n_k/n)$. As a shorthand, when discussing the information in a random variable X that is distributed according to P, we also write $H[X]$. Thus, to the extent that $H[X] \leq \log_2 k$, as the series length n grows the observer can effectively compress the original series of observations and so use less storage than $n \log_2 k$.

The relationship between the raw measurement ($\log_2 k$) and the average-case view ($H[X]$), that we just laid out explicitly, is illustrated in the contrast between Figs. 1(a) and 1(b). The difference $R_1 = \log_2 k - H[X]$ is the amount of redundant information in the raw measurements. As such, the magnitude of R_1 indicates how much they can be compressed.

Information storage can be reduced further, since using $H[X]$ as the amount of information in a measurement implicitly assumed the instrument's outputs were statistically independent. And this, as it turns out, leads to $H[X]$ being an overestimate as to the amount of information in X. For general information sources, there are correlations and restrictions between successive measurements that violate this independence assumption and, helpfully, we can use these to further compress *sequences* of measurements—$X_1, X_2, \ldots, X_\ell$. Concretely, information theory tells us that the irreducible information per observation is given by the Shannon *entropy rate*:

$$
h_\mu = \lim_{\ell \to \infty} \frac{H(\ell)}{\ell} \ , \tag{1}
$$

where $H(\ell) = -\sum_{\{x^\ell\}} \Pr(x^\ell) \log_2 \Pr(x^\ell)$ is the *block entropy*—the Shannon entropy of the length-ℓ word distribution $\Pr(x^\ell)$.

The improved view of the information in a measurement is given in Fig. 1(c). Specifically, since $h_\mu \leq H[X]$, we can compress even more; indeed, by an amount $R_\infty = \log_2 k - h_\mu$.

These comments are no more than a review of basic information theory [1] that used a little algebra. They do, however, set the stage for a parallel, but more detailed, analysis of the information in an observation. In focusing on a single measurement, the following complements recent, more sophisticated analyses of information sources that focused on a process's hidden states [2, and references therein]. In the sense that the latter is a state-centric informational analysis of a process, the following takes the complementary measurement-centric view.

Partly as preview and partly to orient ourselves on the path to be followed, we illustrate the main results in a pictorial fashion similar to that just given; see Fig. 2 which further dissects the information in X.

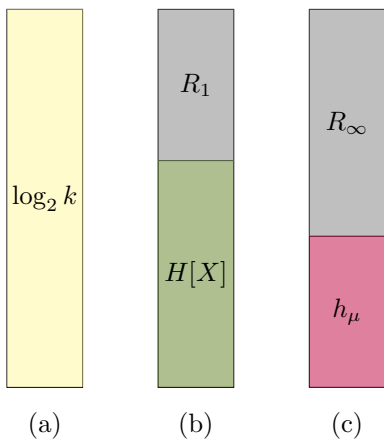As a first cut, the information $H[X]$ provided by each

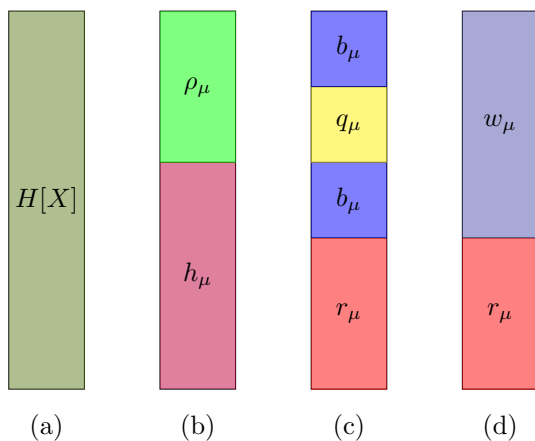FIG. 1. Dissecting information in a single measurement X being one of k values.

FIG. 2. Systematic dissection of $H[X]$.

observation (Fig. 2(a)) can be broken into two pieces: one part is information ρ_μ that could be anticipated from prior observations and the other h_μ — the random component — is that which could not be anticipated. (See Fig. 2(b).) Each of these pieces can be further decomposed into two parts. The random component h_μ breaks into two kinds of randomness: a part b_μ relevant for predicting the future, while the remaining part r_μ is ephemeral, existing only for the moment.

The redundant portion ρ_μ of $H[X]$ in turn splits into two pieces. The first part—also b_μ when the process is stationary—is shared between the past and the current observation, but its relevance stops there. The second piece q_μ is anticipated by the past, is present currently, and also plays a role in future behavior. Notably, this informational piece can be negative. (See Fig. 2(c).)

We can further combine all elements of $H[X]$ that participate in structure — whether it be past, future, or both — into a single element w_μ. This decomposition of $H[X]$ provides a very different decomposition than h_μ and ρ_μ.

It partitions $H[X]$ into a piece w_μ that is structural and a piece r_μ that, as mentioned above, is ephemeral. (See Fig. 2(d).)

With the basic informational components contained in a single measurement laid out, we now derive them from first principles. The next step is to address information in collections of random variables, helpful in a broad array of problems. We then specialize to time series; viz., one-dimensional chains of random variables.

III. INFORMATION MEASURES

Shannon's information theory [1] is a widely used mathematical framework with many advantages in the study of complex, nonlinear systems. Most importantly, it provides a unified quantitative way to analyze systems with broadly dissimilar physical substrates. It further makes no assumptions as to the types of correlation between variables, picking up multi-way nonlinear interactions just as easily as simple pairwise linear correlations.

The workhorse of information theory is the Shannon *entropy* of a random variable, just introduced. The entropy measures what would commonly be considered the amount of information learned, on average, from observing a sample from that random variable. The entropy $H[X]$ of a random variable X taking on values $x \in \mathcal{A} = \{1, \ldots, k\}$ with distribution $\Pr(X = x)$ has the following functional form:

$$H[X] = -\sum_{x \in \mathcal{A}} \Pr(x) \log_2 \Pr(x) . \tag{2}$$

The entropy is defined in the same manner over joint random variables—say, X and Y—where the above distribution is replaced by the joint probability $\Pr(X, Y)$.

When considering more than a single random variable, it is quite reasonable to ask how much uncertainty remains in one variable given knowledge of the other. The average entropy in one variable X given the outcome of another variable Y is the *conditional entropy*:

$$H[X|Y] = H[X, Y] - H[Y] . \tag{3}$$

That is, it is the entropy of the joint random variable (X, Y) with the marginal entropy $H[Y]$ of Y subtracted from it.

The fundamental measure of correlation between random variables is the *mutual information*. As stated before, it can be adapted to measure all kinds of interaction between two variables. It can be written in several forms,

including:

$$I[X;Y] = H[X] + H[Y] - H[X,Y] \qquad (4)$$
$$= H[X,Y] - H[X|Y] - H[Y|X] \; . \qquad (5)$$

Two variables are generally considered *independent* if their mutual information is zero.

Like the entropy, the mutual information can also be conditioned on another variable, say Z, resulting in the *conditional mutual information*. Its definition is a straightforward modification of Eq. (4):

$$I[X;Y|Z] = H[X|Z] + H[Y|Z] - H[X,Y|Z] \; . \qquad (6)$$

For example, consider two random variables X and Y that take the values 0 or 1 independently and uniformly, and a third $Z = X \textbf{ XOR } Y$, the exclusive-or of the two. There is a total of two bits of information among the three variables: $H[X,Y,Z] = 2$ bits. Furthermore, the variables X and Y share a single bit of information with Z, their parity. Thus, $I[X,Y;Z] = 1$ bit. Interestingly, although X and Y are independent, $I[X;Y] = 0$, they are not conditionally independent: $I[X;Y|Z] = 1$.

IV. MULTIVARIATE INFORMATION MEASURES

We now turn to a difficult problem: How does one quantify interactions among an arbitrary set of variables? As just noted, the mutual information provides a very general, widely applicable method of measuring dependence between *two*, possibly composite, random variables. The challenge comes in the fact that there exist several distinct methods for measuring dependence between more than two random variables.

In the following, we will use the word *information* to name measures which involve only subsets of entropy which are shared by multiple variables, and the word *entropy* for those measures which do not. The distinction is made visually clear in Fig. 7. Further, although we are aware of the arbitrariness of the names of these measures and the connotations the names may or may not carry, we do not wish the reader to dwell on them. Rather, we would like to stress that it is the quantities themselves we are interested in, and that the I-diagrams of Sec. VI A provide an unbiased and unencumbered method of understanding them.

Consider a finite set $\mathcal{A}$ and random variables X_i taking on values $x_i \in \mathcal{A}$ for all $i \in \mathbb{Z}$. The vector of N random variables $X_{0:N} = \{X_0, X_1, \ldots, X_{N-1}\}$ takes on values in $\mathcal{A}^N$. A straightforward generalization of Eq. (2) yields

the *joint entropy*:

$$H[X_{0:N}] = - \sum_{\{x_{0:N}\}} \Pr(x_{0:N}) \log_2 \Pr(x_{0:N}) \; , \qquad (7)$$

which measures the total amount of information contained in the joint distribution. By this we mean that if we wanted to convey the particular realization this random variable took, we would need to transmit $H[X_{0:N}]$ bits. From here onward, we suppress notating the set $\{x_{0:N}\}$ of realizations over which the sums are taken.

In generalizing the mutual information to arbitrary sets of variables, we make use of power sets. We let $\Omega_N = \{0, 1, \ldots, N-1\}$ denote the universal set over the variable indices and define $P(N) = \mathcal{P}(\Omega_N)$ as the power set over Ω_N. Then, for any set $A \in P(N)$, its complement is denoted $\bar{A} = \Omega_N \setminus A$ and its cardinality is denoted $|A|$. Finally, we use a shorthand to refer to the set of random variables corresponding to index set A:

$$X_A \equiv \{X_i : i \in A\} \; . \qquad (8)$$

There are at least three extensions of the two-variable mutual information, each based on a different interpretation of what its original definition intended. The first is the *multivariate mutual information* or *co-information* [3]: $I[X_0; X_1; \ldots; X_{N-1}]$. Denoted $I[X_{0:N}]$, it is the amount of mutual information to which *all* variables contribute:

$$I[X_{0:N}] = - \sum \Pr(x_{0:N}) \log_2 \left(\prod_{A \in P(N)} \Pr(x_A)^{-1^{|A|}} \right)$$
$$= - \sum_{A \in P(N)} (-1)^{|A|} H[X_A] \qquad (9)$$
$$= H[X_{0:N}] - \sum_{\substack{A \in P(N) \\ 0 < |A| < N}} I[X_A | X_{\bar{A}}] \; , \qquad (10)$$

where, e.g., $I[X_{\{1,3,4\}} | X_{\{0,2\}}] = I[X_1; X_3; X_4 | X_0, X_2]$ and $I[X_{\{1\}} | X_{\{0,2\}}] = H[X_1 | X_0, X_2]$. It can be verified that Eq. (9) is a generalization of Eq. (4), adding and subtracting all possible entropies according to the number of random variables they include. We will now demonstrate the usage of our notation by way of an example, first using Eq. 9:

$$I[X_0; X_1; X_2] = H[X_0] + H[X_1] + H[X_2]$$
$$- H[X_0, X_1] - H[X_0, X_2] - H[X_1, X_2]$$
$$+ H[X_0, X_1, X_2]$$

and using Eq. 10:

$$
\begin{aligned}
I[X_0; X_1; X_2] = {} & H[X_0, X_1, X_2] - H[X_0|X_1, X_2] \\
& - H[X_1|X_0, X_2] - H[X_2|X_0, X_1] \\
& - I[X_0; X_1|X_2] - I[X_0; X_2|X_1] \\
& - I[X_1; X_2|X_0]
\end{aligned}
$$

The multivariate mutual information has several interesting properties. First, it can be negative, though a consistent interpretation of what this means is still lacking in the literature. Second, this measure vanishes if *any two* variables in the set are completely independent. (That is, they are independent and also conditionally independent with respect to all subsets of the other variables.) This is true regardless of interdependencies among the other variables. The multivariate mutual information has been used in the study of gene-environment interactions [4].

In the second interpretation, the mutual information is seen as the relative entropy between a joint distribution and the product of its marginals. Specifically, the starting point is:

$$
I[X; Y] = \sum \Pr(x, y) \log_2 \frac{\Pr(x, y)}{\Pr(x) \Pr(y)} , \qquad (11)
$$

which is simply a rewriting of Eq. (4). When generalized from this form, we obtain the *total correlation* [5]:

$$
\begin{aligned}
T[X_{0:N}] &= \sum \Pr(x_{0:N}) \log_2 \left(\frac{\Pr(x_{0:N})}{\Pr(x_0) \dots \Pr(x_N)} \right) \\
&= \sum_{\substack{A \in P(N) \\ |A|=1}} H[X_A] - H[X_{0:N}] . \qquad (12)
\end{aligned}
$$

The total correlation is sometimes referred to as the "multi-information" or the "integration", though we refrain from using these ambiguous terms. It differs from the prior measure in many fundamental ways. To begin with, it is nonnegative. It also differs in that if X_0 is independent of the others, then $T[X_{0:N}] = T[X_{1:N}]$. Finally, it compares only the individual variables to the entire set. This leads it to include redundant correlations in the set, giving a possibly misleading representation of the dependencies contained in the system. Indeed, this is a common problem. The total correlation and the net measure miss, or at best conflate unique $(n > 2)$-way interactions. The total correlation has been used to analyze frequency data [6].

The last extension stems from the view that mutual information is the joint entropy minus all (single-variable) unshared information—that is, we start from Eq. (5). When interpreted this way, the generalization is called the *binding information* [7]:

$$
B[X_{0:N}] = H[X_{0:N}] - \sum_{\substack{A \in P(N) \\ |A|=1}} H[X_A|X_{\bar{A}}] . \qquad (13)
$$

Like the total correlation, the binding information is nonnegative and independent random variables do not change its value. Note that $B[X_{0:N}]$ is a first approximation to the multivariate information of Eq. (9) when the sets A are restricted to singleton sets. Binding information was derived in the process of defining a possible measure of musical "interestingness" [8].

We next define three additional multivariate information measures that have not been studied previously, but appear following a similar strategy. First, we have the amount of information in individual variables that is not shared in any way. This is the *residual entropy*:

$$
\begin{aligned}
R[X_{0:N}] &= H[X_{0:N}] - B[X_{0:N}] \\
&= \sum_{\substack{A \in P(N) \\ |A|=1}} H[X_A|X_{\bar{A}}] . \qquad (14)
\end{aligned}
$$

In a sense, it is an anti-mutual information: It measures the total amount of randomness localized to an individual variable and so not correlated to that in its peers.

Second, we can sum the total correlation and the binding information. Then we have the *local exogenous information*:

$$
\begin{aligned}
W[X_{0:N}] &= B[X_{0:N}] + T[X_{0:N}] \qquad (15) \\
&= \sum_{\substack{A \in P(N) \\ |A|=1}} (H[X_A] - H[X_A|X_{\bar{A}}]) \qquad (16) \\
&= \sum_{\substack{A \in P(N) \\ |A|=1}} I[X_A; X_{\bar{A}}] . \qquad (17)
\end{aligned}
$$

It is the amount of information in each variable that comes from its peers. It is a "very mutual" information, one that discounts for the randomness produced locally—that randomness inherent in each variable individually.

$W[X_{0:N}]$ is close to the binding information, except that it uses the sum of marginals not the joint entropy. As such, it seems to more consistently capture the role of single variables within a set than $B[X_{0:N}]$, which compares the set's joint entropy to individual residual uncertainties.

Third and finally, there is a measure which, for lack of a better name, we call the *enigmatic information*:

$$
Q[X_{0:N}] = T[X_{0:N}] - B[X_{0:N}] . \qquad (18)
$$

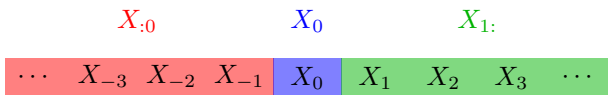Like the multivariate mutual information—which it

FIG. 3. A process's time series: Time indices less than zero refer to the past $X_{:0}$; index 0 to the present X_0; and times after 0 to the future $X_{1:}$.

equals when $N = 3$—it can be negative. Its operational meaning will become clear on further discussion.

Finally, there are a number of connections between these multivariate informations and other formulations. We have in mind, in particular, connections to the entropy expansions of Ref. [9]. We draw these out explicitly in a sequel.

V. TIME SERIES

We now adapt the general multivariate measures to analyze discrete-valued, discrete-time series generated by a stationary process. That is, rather than analyzing sets of random variables, we specialize to a one-dimensional chain of them. In this setting, the measures are most appropriately applied to successively longer blocks of consecutive observations. This allows us to study the asymptotic block-length behavior of each, mimicking the approach of Ref. [2, 10]. For the class of processes known as *finitary* (defined shortly), each of these measures tend to a linear asymptote characterized by a subextensive component and an extensive component controlled by an asymptotic growth rate.

Let's first state more precisely and introduce the notation for the class of processes that are the object of study. We consider a bi-infinite chain $\ldots X_{-1} X_0 X_1 \ldots$ of random variables. Each $X_t, t \in \mathbb{Z}$, takes on a finite set of values $x_t \in \mathcal{A}$. We denote contiguous subsets of the time series with $X_{A:B}$ where the left index is inclusive and the right is exclusive. By leaving one of the indices off the subset is partially infinite in that direction. We divide this bi-infinite chain into three segments. First we single out the *present* X_0. All the symbols prior to the present are the *past* $X_{:0}$. The symbols following the present are the *future* $X_{1:}$. Figure 3 illustrates the setting.

Our focus is on the ℓ-blocks $X_{t:t+\ell} = X_t X_{t+1} \cdots X_{t+\ell-1}$. The associated *process* is specified by the set of length-ℓ word distributions: $\{\Pr(X_{t:t+\ell}) : t \in \mathbb{Z}, \ell \in \mathbb{N}\}$. We consider only *stationary* processes for which $\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell})$. And so, we drop the absolute-time index t. More precisely, the word probabilities derive from an underlying time-shift invariant, ergodic measure μ on the space of bi-infinite sequences.

In the following, an information measure $\mathcal{F}$ applied to the process's length-ℓ words is denoted $\mathcal{F}[X_{0:\ell}]$ or, as a shorthand, $\mathcal{F}(\ell)$.

A. Block Entropy versus Total Correlation

We begin with the long-studied block entropy information measure $H(\ell)$ [11, 12]. (For a review and background to the following see Ref. [10].) The block entropy curve defines two primary features. First, its growth rate limits to the *entropy rate* h_μ. Second, its subextensive component is the *excess entropy* $\mathbf{E}$:

$$\mathbf{E} = I[X_{:0}; X_{0:}] , \qquad (19)$$

which expresses the totality of information shared between the past and future.

The entropy rate and excess entropy, and the way in which they are approached with increasing block length, are commonly used quantifiers for complexity in many fields. They are complementary in the sense that, for finitary processes, the block entropy for sufficiently long blocks takes the form:

$$H(\ell) \sim \mathbf{E} + \ell h_\mu . \qquad (20)$$

Recall that $H(0) = 0$ and that $H(\ell)$ is monotone increasing and concave down. The finitary processes, mentioned above, are those with finite $\mathbf{E}$.

Next, we turn to a less well studied measure for time series—the *block total correlation* $T(\ell)$. Adapting Eq. (12) to a stationary process gives its definition:

$$T(\ell) = \ell H[X_0] - H(\ell) . \qquad (21)$$

Note that $T(0) = 0$ and $T(1) = 0$. Effectively, it compares a process's block entropy to the case of independent, identically distributed random variables. In many ways, the block total correlation is the reverse side of an information-theoretic coin for which the block entropy is the obverse. For finitary processes, its growth rate limits to a constant ρ_μ and its subextensive part is a constant that turns out to be $-\mathbf{E}$:

$$T(\ell) \sim -\mathbf{E} + \ell \rho_\mu . \qquad (22)$$

That is, $\rho_\mu = \lim_{\ell \to \infty} T(\ell)/\ell$. Finally, $T(\ell)$ is monotone increasing, but concave up. All of this is derived directly from Eqs. (20) and (21), by using well known properties of the block entropy.

The block entropy and block total correlation are plotted in Fig. 4. Both measures are 0 at $\ell = 0$ and from there approach their asymptotic behavior, denoted by
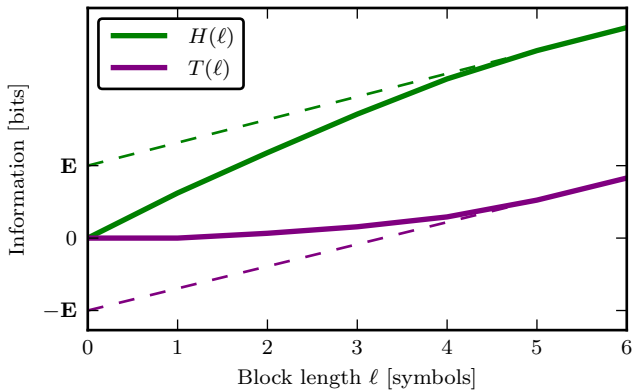
FIG. 4. Block entropy $H(\ell)$ and block total correlation $T(\ell)$ illustrating their behaviors for the NRPS Process.

the dashed lines. Though their asymptotic slopes appear to be the same, they in fact differ. Numerical data for the asymptotic values can be found in Tables I and II under the heading NRPS (defined later).

There is a persistent confusion in the neuroscience, complex systems, and information theory literatures concerning the relationship between block entropy and block total correlation [13]. This can be alleviated by explicitly demonstrating a partial symmetry between the two in the time series setting and by highlighting a weakness of the total correlation.

We begin by showing how, for stationary processes, the block entropy and the block total correlation contain much the same information. From Eqs. (7) and (12) we immediately see that:

$$H(\ell) + T(\ell) = \ell H(1) \ . \qquad (23)$$

Furthermore, by substituting Eqs. (20) and (22) in Eq. (23) we note that the righthand side has no subextensive component. This gives further proof that the subextensive components of Eqs. (20) and (22) must be equal and opposite, as claimed. Moreover, by equating individual ℓ-terms we find:

$$h_\mu + \rho_\mu = H(1) \ . \qquad (24)$$

And, this is the decomposition given in Fig. 2(b): the lefthand side provides two pieces comprising the single-observation entropy $H(1)$.

Continuing, either information measure can be used to obtain the excess entropy. In addition, since the block entropy provides h_μ as well as intrinsically containing $H(1)$, ρ_μ can be directly obtained from the block entropy function by taking $H(1) - h_\mu$, yielding ρ_μ. The same is not true, however, for the total correlation. Though ρ_μ can be computed, one cannot obtain h_μ from $T(\ell)$ alone—

$H(1)$ is required, but not available from $T(\ell)$, since it is subtracted out.

There are further parallels between the two quantities that can be drawn. First, following Ref. [10], we define discrete derivatives of the block measures at length ℓ:

$$h_\ell = H(\ell) - H(\ell - 1) \qquad (25)$$
$$\rho_\ell = T(\ell) - T(\ell - 1) \ . \qquad (26)$$

These approach h_μ and ρ_μ, respectively. From them we can determine the subextensive components by discrete integration, while subtracting out the asymptotic behavior. We find that:

$$\mathbf{E} = \sum_{\ell=1}^{\infty} (h_\ell - h_\mu) \qquad (27)$$

and also that

$$\mathbf{E} = -\sum_{\ell=1}^{\infty} (\rho_\ell - \rho_\mu) \ . \qquad (28)$$

Second, these sums are equal term by term.

The first sum, however, indirectly brings us back to Eq. (24). Since $h_1 = H(1)$, we have:

$$\mathbf{E} = \rho_\mu + \sum_{\ell=2}^{\infty} (h_\ell - h_\mu) \ . \qquad (29)$$

Finally, it has been said that the total correlation ("multi-information") is the first term in $\mathbf{E}$ [14]. This has perhaps given the impression that the total correlation is only useful as a crude approximation. Equation (29) shows that it is actually the total correlation rate ρ_μ that is $\mathbf{E}$'s first term. As we just showed, the total correlation is more useful than being a first term in an expansion. Its utility is ultimately limited, though, since its properties are redundant with that of the block entropy which, in addition, gives the process's entropy rate h_μ.

B. A Finer Decomposition

We now show how, in the time series setting, the binding information, local exogenous information, enigmatic information, and residual entropy constitute a refinement of the single-measurement decomposition provided by the block entropy and the total correlation [7, 8]. To begin,

their block equivalents are, respectively:

$$B(\ell) = H(\ell) - R(\ell) \tag{30}$$
$$Q(\ell) = T(\ell) - B(\ell) \tag{31}$$
$$W(\ell) = B(\ell) + T(\ell) \ , \tag{32}$$

where $R(\ell)$ does not have an analogously simple form. Their asymptotic behaviors are, respectively:

$$R(\ell) \sim \mathbf{E}_R + \ell r_\mu \tag{33}$$
$$B(\ell) \sim \mathbf{E}_B + \ell b_\mu \tag{34}$$
$$Q(\ell) \sim \mathbf{E}_Q + \ell q_\mu \tag{35}$$
$$W(\ell) \sim \mathbf{E}_W + \ell w_\mu \ . \tag{36}$$

Their associated rates break the prior two components (h_μ and ρ_μ) into finer pieces. Substituting their definitions into Eqs. (7) and (21) we have:

$$H(\ell) = B(\ell) + R(\ell) \tag{37}$$
$$= (\mathbf{E}_B + \mathbf{E}_R) + \ell(b_\mu + r_\mu) \tag{38}$$
$$T(\ell) = B(\ell) + Q(\ell) \tag{39}$$
$$= (\mathbf{E}_B + \mathbf{E}_Q) + \ell(b_\mu + q_\mu) \ . \tag{40}$$

The rates in Eqs. (38) and (40) corresponding to h_μ and ρ_μ, respectively, give the decomposition laid out in Fig. 2(c) above. Two of these components (b_μ and r_μ) were defined in Ref. [7] and the third (q_μ) is a direct extension. We defer interpreting them to Sec. VI B which provides greater understanding by appealing to the semantics afforded by the process information diagram developed there.

The local exogenous information, rather than refining the decomposition provided by the block entropy and the total correlation, provides a different decomposition:

$$W(\ell) = B(\ell) + T(\ell) \tag{41}$$
$$= (\mathbf{E}_B - \mathbf{E}) + \ell(b_\mu + \rho_\mu) \ . \tag{42}$$

So, $w_\mu = h_\mu + \rho_\mu$, as mentioned in Fig. 2(d).

Similar to Eq. (23), we can take the local exogenous information together with the residual entropy and find:

$$R(\ell) + W(\ell) = \ell H(1) \ . \tag{43}$$

This implies that $\mathbf{E}_R = -\mathbf{E}_W$ and that r_μ and w_μ are yet another partitioning of $H[X]$, as shown earlier in Fig. 2(d).

Figure 5 illustrates these four block measures for a generic process. Each of the four measures reaches asymptotic linear behavior at a length of $\ell = 9$ symbols. Once there, we see that they each possess a slope that we just showed to be a decomposition of the slopes from the
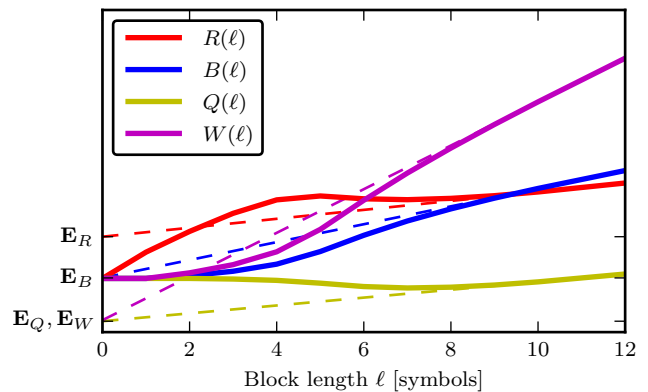

FIG. 5. Block equivalents of the residual entropy $R(\ell)$, binding information $B(\ell)$, enigmatic information $Q(\ell)$, and local exogenous information $W(\ell)$ for a generic process (same as previous figure).

measures in Fig. 4. Furthermore, each has a subextensive component that is found as the y-intercept of the linear asymptote. These subextensive parts provide a decomposition of the excess entropy, discussed further below in Sec. VI B 3.

C. Multivariate Mutual Information

Lastly, we come to the block equivalent of the multivariate mutual information $I[X_{0:N}]$:

$$I(\ell) = H(\ell) - \sum_{\substack{A \in P(\ell) \\ 0 < |A| < \ell}} I[X_A | X_{\bar{A}}] \ . \tag{44}$$

Superficially, it scales similarly to the other measures:

$$I(\ell) \sim \mathbf{I} + \ell i_\mu \ , \tag{45}$$

with an asymptotic growth rate i_μ and a constant subextensive component $\mathbf{I}$. Yet, it has differing implications regarding what it captures in the process. This is drawn out by the following propositions, whose proofs appear elsewhere.

The first concerns the subextensive part of $I(\ell)$.

Proposition 1. *For all finite-state processes:*

$$h_\mu > 0 \quad \Rightarrow \quad \lim_{\ell \to \infty} I(\ell) = 0 \ . \tag{46}$$

The intuition behind this is fairly straightforward. For $I(\ell)$ to be nonzero, no two observations can be independent. Finite-state processes with positive h_μ are stochastic, however. So, observations become (conditionally) decoupled exponentially fast. Thus, for arbitrarily long
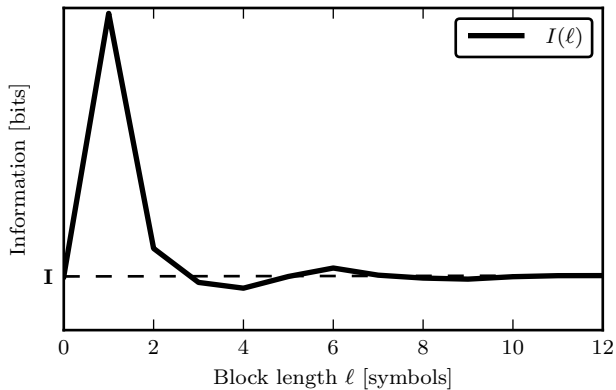
FIG. 6. Block multivariate mutual information $I(\ell)$ for the same example process as before.

blocks, the first and the last observations tend toward independence exponentially and so $I(\ell)$ limits to 0.

The second proposition regards the growth rate i_μ.

Proposition 2. *For all finite-state processes:*

$$i_\mu = 0 \ . \tag{47}$$

The intuition behind this follows from the first proposition. If $h_\mu > 0$, then it is clear that since $I(\ell)$ tends toward 0, then the slope must also tend toward 0. What remains are those processes that are finite state but for which $h_\mu = 0$. These are the periodic processes. For them, i_μ also vanishes since, although $I(\ell)$ may be nonzero, there is a finite amount of information contained in a bi-infinite periodic sequence. Once all this information has been accounted for at a particular block length, then for all blocks larger than this there is no additional information to gain. And so, i_μ decays to 0.

The final result concerns the subextensive component **I**.

Proposition 3. *For all finite-state processes with* $h_\mu > 0$:

$$\mathbf{I} = 0 \ . \tag{48}$$

This follows directly from the previous two propositions.

Thus, the block multivariate mutual information is qualitatively different from the other block measures. It appears to be most interesting for infinitary processes with infinite excess entropy.

Figure 6 demonstrates the general behavior of $I(\ell)$, illustrating the three propositions. The dashed line highlights the asymptotic behavior of $I(\ell)$: both **I** and i_μ

vanish. We further see that $I(\ell)$ is not restricted to positive values. It oscillates about 0 until length $\ell = 11$ where it finally vanishes.

VI. INFORMATION DIAGRAMS

Information diagrams [15] provide a graphical and intuitive way to interpret the information-theoretic relationships among variables. In construction and concept, they are very similar to Venn diagrams. The key difference is that the measure used is a Shannon entropy rather than a set size. Additionally, an overlap is not set intersection but rather a mutual information. The irreducible intersections are, in fact, elementary *atoms* of a sigma-algebra over the random-variable event space. An atom's size reflects the magnitude of one or another *Shannon information measure*—marginal, joint, or conditional entropy or mutual information.

A. Four-Variable Information Diagrams

Using information diagrams we can deepen our understanding of the multivariate informations defined in Sec. IV. Fig. 7 illustrates them for four random variables—X, Y, Z, W. There, an atom's shade of gray denotes how much weight it carries in the overall value of its measure. Consider for example the total correlation I-diagram in Fig. 7(c). From the definition of the total correlation, Eq. (12), we see that each variable provides one count to each of its atoms and then a count is removed from each atom. Thus, the atom associated with four-way intersection $W \cap X \cap Y \cap Z$ contained in each of the four variables carries a total weight $I[W; X; Y; Z] = 4 - 1 = 3$. Those atoms contained in three variables carry a weight of 2, those shared among only two variables a weight of 1, and information solely contained in one variable is not counted at all.

Utilizing the I-diagrams in Fig. 7, we can easily visualize and intuit how these various information measures relate to each other and the distributions they represent. In Fig. 7(a), we find the joint entropy. Since it represents all information contained in the distribution with no bias to any sort of interaction, we see that it counts each and every atom once. The residual entropy, Fig. 7(e), is equally easy to interpret: it counts each atom which is not shared by two or more variables.

The distinctions in the menagerie of measures attempting to capture interactions among N variables can also be easily seen. The multivariate mutual information, Fig. 7(b), stands out in that it is isolated to a single atom, that contained in all variables. This makes it

(a) Joint entropy

(b) Multivariate mutual information (c) Total correlation

(d) Binding informa- (e) Residual entropy
tion

(f) Local exogenous (g) Enigmatic infor-
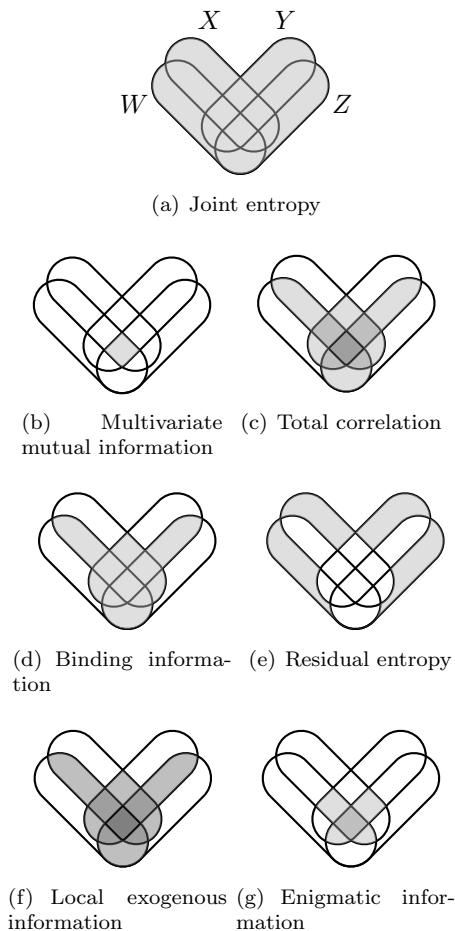information   mation

FIG. 7. Four-variable information diagrams for the multivariate information measures of Sec. IV. Darker shades of gray denote heavier weighting in the corresponding informational sum. For example, the atoms to which all four variables contribute are added thrice to the total correlation and so the central atom's weight $I[W; X; Y; Z] = 3$.

clear why the independence of any two of the variables leads to a zero value for this measure. The total correlation, Fig. 7(c), contains all atoms contained in at least two variables and gives higher weight to those contained in more variables. The local exogenous information, Fig. 7(f), is similar. It counts the same atoms as the total correlation does, but it gives them higher weight. Lastly, the binding information, Fig. 7(d), also counts the same atoms, but only weights each of them once regardless of how many variables they participate in.

The lone enigmatic information, Fig. 7(g), counts only those variables that participate in at least three variables and, similar to the total correlation, it counts those that participate in more variables more heavily.

B. Process Information Diagrams

Following Ref. [16] we adapt the multivariate I-diagrams just laid out to tracking information in finitary stationary processes. In particular, we develop process I-diagrams to explain the information in a single observation, as described before in Fig. 2. The resulting process I-diagram is displayed in Fig. 8. As we will see, exploring the diagram gives a greater, semantic understanding of the relationships among the process variables and, as we will emphasize, of the internal structure of the process itself.

For all measures, except the multivariate mutual information, the extensive rate corresponds to one or more atoms in the decomposition of $H[X_0]$. To begin, we allow $H[X_0]$ to be split in two by the past. This exposes two pieces: h_μ, the part exterior to the past, and ρ_μ, the part interior. This partitioning has been well studied in information theory due to how it naturally arises as one observes a sequence. This decomposition is displayed in Fig. 9(a).

Taking a step back and including the future in the diagram, we obtain a more detailed understanding of how information is transmitted in a process. The past and the future together divide $H[X_0]$ into four parts; see Fig. 9(b). We will discuss each part shortly. First, however, we draw out a different decomposition—that into r_μ and w_μ as seen in Fig. 9(c). From this diagram it is easy to see the semantic meaning behind the decomposition: r_μ being divorced from any temporal structure, while w_μ is steeped in it.

We finally turn to the partitioning shown in Fig. 9(b). The process I-diagram makes it rather transparent in which sense r_μ is an amount of *ephemeral* information: its atom lies outside both the past and future sets and so it exists only in the present moment, having no repercussions for the future and being no consequence of the past. It is the amount of information in the present observation neither communicated to the future nor from the past. Ref. [7] referred to this as the residual entropy *rate*, as it is the amount of uncertainty that remains in the present even after accounting for every other variable in the time series.

Ref. [7] also proposed to use b_μ as a measure of structural complexity [7], and we tend to agree. The argument for this is intuitive: b_μ is an amount of information that is present now, is not explained by the past, but has repercussions in the future. That is, it is the portion of the entropy rate h_μ that has consequences. In some contexts one may prefer to employ the ratio b_μ/h_μ when b_μ is interpreted an indicator of complex behavior since, for a fixed b_μ, larger h_μ values imply less temporal structure in the time series.
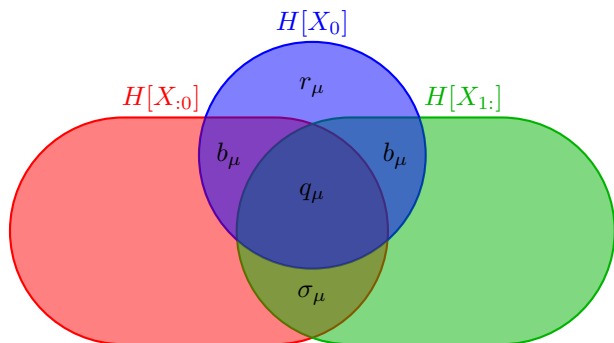
FIG. 8. I-diagram anatomy of $H[X_0]$ in the full context of time: The past $X_{:0}$ partitions $H[X_0]$ into two pieces: h_μ and ρ_μ. The future $X_{0:}$ then partitions those further into r_μ, two b_μs, and q_μ. This leaves a component σ_μ, shared by the past and the future, that is not in the present X_0.

FIG. 9. The three decompositions of $H[X]$ from Fig. 2. The dissecting lines are identical to those in Fig. 8.

Due to stationarity, the mutual information $I[X_0; X_{1:}|X_{:0}]$ between the present X_0 and the future $X_{1:}$ conditioned on the past $X_{:0}$ is the same as the mutual information $I[X_0; X_{:0}|X_{1:}]$ between X_0 and the past $X_{:0}$ conditioned on the future $X_{1:}$. Moreover, both are b_μ. This lends a symmetry to the process I-diagram that does not exist for nonstationary processes. Thus, b_μ atoms in Fig. 8 are the same size.

There are two atoms remaining in the process I-diagram that have not been discussed in literature. Both merit attention. The first is q_μ—the information shared by the past, the present, and the future. Notably, its value can be negative and we discuss this further below in Sec. VI B 1. The other piece, denoted σ_μ, is a component of information shared between the past and the future that *does not exist in the present* observation. This piece is vital evidence that attempting to understand a process without using a model for its generating mechanism is ultimately incomplete. We discuss this point further in Sec. VI B 2 below.

1. Negativity of q_μ

The sign of q_μ holds valuable information. To see what this is we apply the partial information decomposition [17] to further analyze $w_\mu = I[X_0; X_{:0}, X_{1:}]$—that portion of the present shared with the past and future. By decomposing w_μ into four pieces—three of which are unique—we gain greater insight into the value of q_μ and also draw out potential asymmetries between the past and the future.

The partial information lattice provides us with a method to isolate (i) the contributions $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ to w_μ that both the past and the future provide redundantly, (ii) parts $\Pi_{\{X_{:0}\}}$ and $\Pi_{\{X_{1:}\}}$ that are uniquely
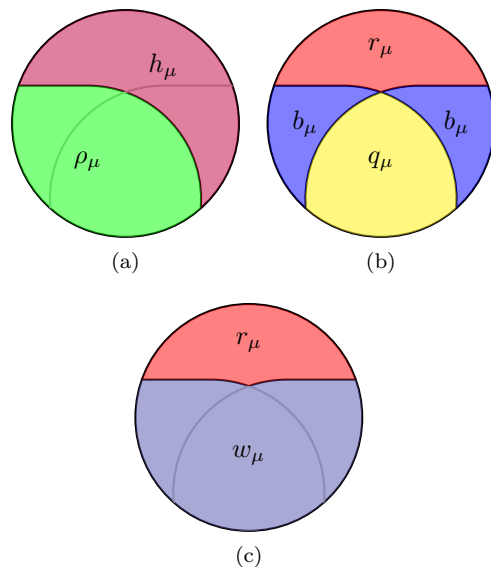
provided by the past and the future, respectively, and (iii) a part $\Pi_{\{X_{:0}, X_{1:}\}}$ that is synergistically provided by both the past and the future. Note that, due to stationarity, $\Pi_{\{X_{:0}\}} = \Pi_{\{X_{1:}\}}$. We refer to this as the *uniquity* and denote it ι.

Using Ref. [17] we see that q_μ is equal to the *redundancy* minus the *synergy* of the past and the future, when determining the present. Thus, if $q_\mu > 0$, the past and future predominantly contribute information to the present. When $q_\mu < 0$, however, considering the past and the future separately in determining the present misses essential correlations. The latter can be teased out if the past and future are considered together.

The process I-diagram (Fig. 8) showed that the mutual information between the present and either the past or the future is ρ_μ. One might suspect from this that the past and the future provide the same information to the present, but this would be incorrect. Though they provide the same *quantity* of information to the present, *what* that information conveys can differ. This is evidence of a process's structural irreversibility; cf. Refs. [16, 18]. In this light, the redundancy $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ between the past and future when considering the present is $\rho_\mu - \iota$. Furthermore, the synergy $\Pi_{\{X_{:0}, X_{1:}\}}$ provided by the past and the future is equal to $b_\mu - \iota$.

Taking this all together, we find what we already knew: that $q_\mu = \rho_\mu - b_\mu$, The journey to this conclusion, however, provided us with deeper insight into what negative q_μ means and into the structure of w_μ and the process as a whole.
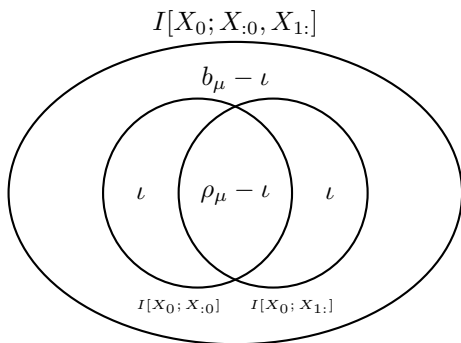
$$I[X_0; X_{:0}, X_{1:}]$$

FIG. 10. Partial information decomposition of $w_\mu = I[X_0; X_{:0}, X_{1:}]$. The multivariate mutual information q_μ is given by the redundancy $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ minus the synergy $\Pi_{\{X_{:0}, X_{1:}\}}$. $w_\mu = \rho_\mu + b_\mu$ is the sum of all atoms in this diagram.

2. Consequence of σ_μ: Why we model

Notably, the final piece of the process I-diagram is not part of $H[X_0]$—not a component of the information in a single observation. This is σ_μ, which represents information that is transmitted from the past to the future, but does not go through the currently observed symbol X_0. This is readily understood and leads to an important conclusion.

If one believes that the process under study is generated according to the laws of physics, then the process's internal physical configuration must store all the information from the past that is relevant for generating future behavior. Only when the observed process is order-1 Markov is it sufficient to keep track of just the current observable. For the plethora of processes that are not order-1 or that are non-Markovian altogether, we are faced with the fact that information relevant for future behavior must be stored somehow. And, this fact is reflected in the existence of σ_μ. When $\sigma_\mu > 0$, a complete description of the process requires accounting for this internal configurational or, simply, *state information*. This is why we build models and cannot rely on only collecting observation sequences.

The amount of information shared between $X_{:0}$ and $X_{1:}$, but ignoring X_0, was previously discussed in Ref. [19]. We now see that the meaning of this information quantity — there denoted $\mathcal{I}_1$ — is easily gleaned from its components: $\mathcal{I}_1 = q_\mu + \sigma_\mu$.

Furthermore, in Refs. [7], [8], and [19], efficient computation of b_μ and $\mathcal{I}_1$ were not provided and the brute force estimates are inaccurate and very compute intensive. Fortunately, by a direct extension of the methods developed in Ref. [18] on bidirectional machines, we can easily compute both $r_\mu = H[X_0|\mathcal{S}_0^+, \mathcal{S}_1^-]$ and $\mathcal{I}_1 = I[\mathcal{S}_0^+, \mathcal{S}_1^-]$. This is done by constructing joint prob-

abilities of forward-time and reverse-time causal states — $\{\mathcal{S}^+\}$ and $\{\mathcal{S}^-\}$, respectively — at different time indices employing the dynamic of the bidirectional machine. This gives closed-form, exact methods of calculating these two measures, provided one constructs the process's forward and reverse ϵ-machines. b_μ follows directly in this case since it is the difference of h_μ and r_μ; the former is also directly calculated from the ϵ-machine.

3. Decompositions of **E**

Using the process I-diagram and the tools provided above, three unique decompositions of the excess entropy, Eq. (19), can be given. Each provides a different interpretation of how information is transmitted from the past to the future.

The first is provided by Eqs. (37)-(40). The subextensive parts of the block entropy and total correlation there determine the excess entropy decomposition. We have:

$$\mathbf{E} = \mathbf{E}_B + \mathbf{E}_R \tag{49}$$
$$= -\mathbf{E}_B - \mathbf{E}_Q \tag{50}$$
$$= \frac{1}{2}(\mathbf{E}_R - \mathbf{E}_Q) \tag{51}$$
$$= -\frac{1}{2}(\mathbf{E}_W + \mathbf{E}_Q) . \tag{52}$$

We leave the meaning behind these decompositions as an open problem, but do note that they are distinct from those discussed next.

The second and third decompositions both derive directly from the process I-diagram of Fig. 8. Without further work, one can easily see that the excess entropy breaks into three pieces, all previously discussed:

$$\mathbf{E} = b_\mu + q_\mu + \sigma_\mu . \tag{53}$$

And, finally, one can perform the partial information decomposition on the mutual information $I[X_{:0}; X_0, X_{1:}]$. The result gives an improved understanding of (i) how much information is uniquely shared with the either the immediate or the more distant future and (ii) how much is redundantly or synergistically with both.

The decompositions provided by the atoms of the process I-diagram and those provided by the subextensive rates of block-information curves are conceptually quite different. It has been shown [10] that the subextensive part of the block entropy and the mutual information between the past and the future, though equal for one dimensional processes, differ in two dimensions. We believe the semantic differences shown here are evidence that the degeneracy of alternate **E**-decompositions breaks in higher dimensions.
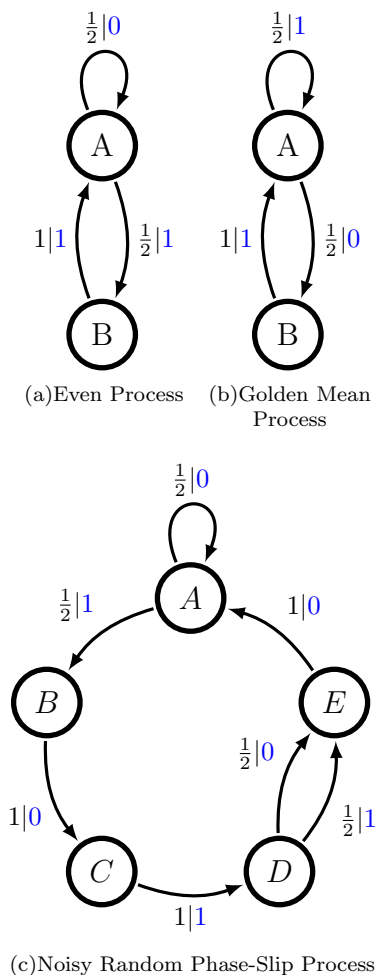
(a)Even Process   (b)Golden Mean Process

(c)Noisy Random Phase-Slip Process

FIG. 11. ϵ-Machine presentations for the three example processes.

VII. EXAMPLES

We now make the preceding concrete by calculating these quantities for three different processes, selected to illustrate a variety of informational properties. Figure 11 gives each process via it's ϵ-machine [20]: the Even Process, the Golden Mean Process, and the Noisy Random Phase-Slip (NRPS) Process. A process's ϵ-machine consists of its *causal states*—a partitioning of infinite pasts into sets that give rise to the same predictions about future behavior. The state transitions are labeled $p|s$ where s is the observed symbol and p is the conditional probability of observing that symbol given the state the process is in. The ϵ-machine representation for a process is its minimal unifilar presentation.

Table I begins by showing the single-observation entropy $H[1]$ followed by h_μ and ρ_μ. Note that the Even and the Golden Mean Processes cannot be differentiated using these measures alone. The table then follows with the finer decomposition. We now see that the processes

can be differentiated. We can understand fairly easily that the Even Process, being infinite-order Markovian, and consisting of blocks of 1s of even length separated by one or more 0s, exhibits more structure than the Golden Mean Process. (This is rather intuitive if one recalls that the Golden Mean Process has only a single restriction: it cannot generate sequences with consecutive 0s.) We see that, for the Even Process, r_μ is 0. This can be understood by considering a bi-infinite sample from the Even Process with a single gap in it. The structure of this process is such that we can always and immediately identify what that missing symbol must be.

These two processes are further differentiated by q_μ, where it is negative for the Even Process and positive for the Golden Mean Process. On the one hand, this implies that there is a larger amount of synergy than redundancy in the Even Process. Indeed, it is often the case, when appealing only to the past or the future, that one cannot determine the value of X_0, but when taken together the possibilities are limited to a single symbol. On the other hand, since q_μ is positive for the Golden Mean Process we can determine that its behavior is dominated by redundant contributions. That w_μ is larger for the Even Process than the Golden Mean Process is consonant with the impression that the former is, overall, more structured.

The next value in the table is σ_μ, the amount of state information not contained in the current observable. This vanishes for the Golden Mean Process, as it is order-1 Markovian. The Even Process, however, has a significant amount of information stored that is not observable in the present.

Last in the table is a partial information decomposition of $I[X_0; X_{:0}, X_{1:}]$. q_μ is given by $\Pi_{\{X_{:0}\}\{X_{1:}\}} - \Pi_{\{X_{:0}, X_{1:}\}}$. Of note here is that the NRPS process's nonzero uniquity $\iota = 0.02437$. For the Even and Golden Mean Processes it vanishes. That is, in the NRPS Process information is uniquely communicated to the present from the past and an equivalent in magnitude, but different, information is communicated to the future. Thus, the NRPS Process illustrates a subtle asymmetry in statistical structure.

Table II then provides an alternate breakdown of $\mathbf{E}$ for each prototype process. We use this here to only highlight how much the processes differ in character from one another. The consequences of the first decomposition of excess entropy—$\mathbf{E} = b_\mu + q_\mu + \sigma_\mu$—follow directly from the previous table's discussion. The second and third decompositions into $\mathbf{E}_R + \mathbf{E}_B$ and $-\mathbf{E}_B - \mathbf{E}_Q$ vary from one another significantly. The Even Process has much larger values for these pieces than the total $\mathbf{E}$, whereas the NRPS process has two values nearly equal to $\mathbf{E}$ and one very small. The Golden Mean Process falls somewhere between these two.

The final excess entropy breakdown is provided by

	Even	Golden Mean	NRPS
$H[1]$	0.91830	0.91830	0.97987
h_μ	0.66667	0.66667	0.50000
ρ_μ	0.25163	0.25163	0.47987
r_μ	0.00000	0.45915	0.16667
b_μ	0.66667	0.20752	0.33333
q_μ	-0.41504	0.04411	0.14654
w_μ	0.91830	0.45915	0.81320
σ_μ	0.66667	0.00000	1.09407
$\Pi_{\{X_{:0}\}\{X_{1:}\}}$	0.25163	0.25163	0.45550
$\iota : \Pi_{\{X_{:0}\}}, \Pi_{\{X_{1:}\}}$	0.00000	0.00000	0.02437
$\Pi_{\{X_{:0},X_{1:}\}}$	0.66667	0.20752	0.30896

TABLE I. Information measure analysis of three processes.

	Even	Golden Mean	NRPS
$\mathbf{E}$	0.91830	0.25163	1.57393
b_μ	0.66667	0.20752	0.33333
q_μ	-0.41504	0.04411	0.14654
σ_μ	0.66667	0.00000	1.09407
$\mathbf{E}_R$	4.48470	0.41504	1.55445
$\mathbf{E}_B$	-3.56640	-0.16341	0.01948
$\mathbf{E}_Q$	2.64810	-0.08822	-1.59342
$\mathbf{E}_W$	-4.48470	-0.41504	-1.55445
$\Pi_{\{X_0\}\{X_{1:}\}}$	0.25163	0.04411	0.47987
$\Pi_{\{X_0\}}$	0.00000	0.20752	0.00000
$\Pi_{\{X_{1:}\}}$	0.00000	0.00000	0.76073
$\Pi_{\{X_0,X_{1:}\}}$	0.66667	0.00000	0.33333

TABLE II. Alternative decompositions of excess entropy $\mathbf{E}$ for the three prototype processes.

the partial information decomposition of $I[X_{:0}; X_0, X_{1:}]$. Here, we again see differing properties among the three processes. The Even Process consists only of redundancy $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ and synergy $\Pi_{\{X_{:0},X_{1:}\}}$. The Golden Mean Process contains no synergy, a small amount of redundancy, and most of its information sharing is with the present uniquely. The NRPS Process possesses both synergy and redundancy, but also a significant amount of information shared solely with the more distant future.

And, finally, Fig. 12 plots how h_μ partitions into r_μ and b_μ for the Golden Mean family of processes. This family consists of all processes with ϵ-machine structure given in Fig. 11(b), but where the outgoing transition probabilities from state $\mathbf{A}$ are parametrized. We can easily see that for small self-loop transition probabilities, the majority of h_μ is consumed by b_μ. This should be intuitive since, when the self-loop probability is small, the process is nearly periodic and r_μ should be nearly zero. On the other end of the spectrum, when the self-loop probability is large, h_μ is mostly consumed by r_μ. This is again intuitive since observations from that process are dominated by 1s and the occasional 0—which provides all the entropy for h_μ—has no effect on structure.
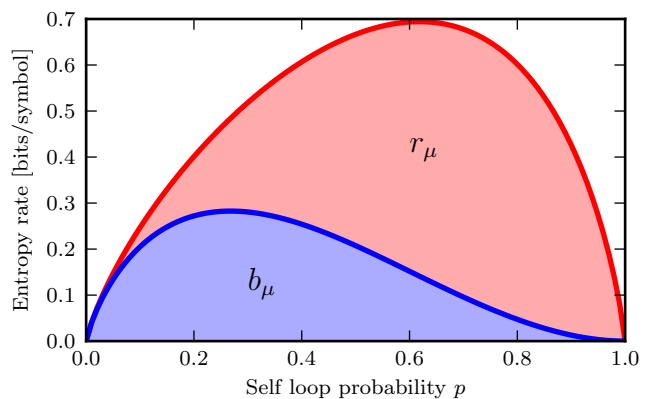
FIG. 12. The breakdown of h_μ for the Golden Mean Process. The self-loop probability was varied from 0 to 1, adjusting the other edge's probability accordingly.

VIII. CONCLUDING REMARKS

We began by outlining a conceptual decomposition of a single observation in a time series: a single observation contains a hierarchy of informational components. We then made the decomposition concrete using a variety of multivariate information measures. Adapting them to time series, we showed that their asymptotic growth rates are identified with the hierarchical decomposition. To unify the various competing views, we provided the measurement-centric process I-diagram, demonstrating that it concisely reveals the semantic meaning behind each component in the hierarchy.

Once the measurement-centric process I-diagram was available, we isolated two components, analyzing in detail their meaning. We utilized the partial information lattice [17] to refine our understanding of when the past and the future redundantly and synergistically inform the present. This allowed us to explain a subtle statistical asymmetry—the directionality in the difference between ρ_μ and $\Pi_{\{X_{:0}\}\{X_{1:}\}}$.

The other atom we singled out in the process I-diagram was σ_μ. It is the most compelling evidence that analyzing a process from its measurements alone, without constructing a state-based model, is ultimately limited. We also quickly stated how a model of the type studied in computational mechanics can be used to give closed-form expressions for many of the quantities discussed which allow for exact calculation of their values.

Next, we discussed how the different methods and measures relate to one of the most widely used complexity measures—the past-future mutual information or excess entropy. In particular, we showed how they yield four distinct decompositions and, in some cases, give useful interpretations of what these decompositions mean operationally.

Then, we calculated all the measures for three different prototype processes, each highlighting particular features of the information-theoretic decompositions. We gave interpretations of negative mutual informations, as seen in q_μ. The interpretations were consistent, understandable, and insightful. In light of the partial information decomposition, there is nothing untoward about negative informations.

By adapting it to the time series setting, we highlighted a key weakness of the total correlation (or multiinformation). This undoubtedly explains the lack of interest in using it in the time series setting, though the weakness still holds when it is used to analyze any group of random variables. The weakness has led to persistent over-interpretations of what it describes. It also may have eclipsed the importance of its more complete analog, such as the block entropy, in the settings of networked random variables.

We believe that the decompositions detailed here provide useful tools for information-theoretic time series analysis, and that r_μ and b_μ potentially have many interesting uses due to their data-centric nature. For example, it is possible that b_μ is useful in constructing errorcorrecting codes since a time series with small $\frac{r_\mu}{h_\mu}$ has a high recovery rate from single-site deletions. This is because if there is no entropy in the value of a site given its neighbors (the past and the future), there is only one value that could have been in that site while still respecting the correlations present. This can be seen clearly with the Even process.

In closing, we take a longer view. There is an exponential number of possible atoms for N-way information measures. In addition, there is a similarly large number possible partial information decompositions for N variables. This diversity presents the possibility of a large number of independent efforts to define and uniquely motivate why one or the other information measure is the best. Indeed, many of these yet-to-be-explored measures may be useful. In this light, there is a bright future for developing information measures adapted to a wide range of nonlinear, complex systems. And, helpfully, a unifying framework appears to be emerging.

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

[2] James P. Crutchfield, Christopher J. Ellison, Ryan G. James, and John R. Mahoney. Synchronization and Control in Intrinsic and Designed Computation: An Information-Theoretic Analysis of Competing Models of Stochastic Computation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(3):037105, July 2010.

[3] Anthony J Bell. The Co-Information Lattice. In N. Murata S. Amari, A. Cichocki , S. Makino, editor, *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*. Springer, New York, ica 2003 edition, 2003.

[4] Pritam Chanda, Aidong Zhang, Daniel Brazeau, Lara Sucheston, Jo L Freudenheim, Christine Ambrosone, and Murali Ramanathan. Information-theoretic metrics for visualizing gene-environment interactions. *American journal of human genetics*, 81(5):939–63, November 2007.

[5] Satosi Watanabe. Information Theoretical Analysis of Multivariate Correlation. *IBM Journal of Research and Development*, 4(1):66–82, January 1960.

[6] Te Sun Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1):26–45, July 1980.

[7] Samer A. Abdallah and Mark D. Plumbley. A Measure of Statistical Complexity Based on Predictive Information. *arXiv:1012.1890v1*, 2010.

[8] Samer A. Abdallah and Mark D. Plumbley. Predictive Information, Multi-Information, and Binding Information. *Technical Report C4DM-TR10-10, Centre for Digital Music, Queen Mary University of London*, 2010.

[9] L. M. A. Bettencourt, V. Gintautas, and M. I. Ham. Identification of functional information subgraphs in complex networks. *Phys. Rev. Lett.*, 100(23):238701, Jun 2008.

[10] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[11] James P. Crutchfield and N. H. Packard. Symbolic Dynamics of Noisy Chaos. *Physica*, 7D:201–223, 1983.

[12] Karl-Erik Eriksson and Kristian Lindgren. Structural Information in Self-Organizing Systems. *Physica Scripta*, 35(3):388–397, March 1987.

[13] For example, see the conflation of I for both the total correlation (here, T) and its rate (here, ρ_μ) in Ref. [14].

[14] Ionas Erb and Nihat Ay. Multi-Information in the Thermodynamic Limit. *Journal of Statistical Physics*, 115(3/4):949–976, May 2004.

[15] Raymond W. Yeung. A New Outlook on Shannon's Information Measures. *IEEE Transactions on Information Theory*, 37(3):466–474, 1991.

[16] James P. Crutchfield, Christopher J. Ellison, and John R. Mahoney. Times Barbed Arrow: Irreversibility, Crypticity, and Stored Information. *Physical Review Letters*, 103(9):094101, 2009.

[17] Paul L. Williams and Randall D. Beer. Non-negative Decomposition of Multivariate Information. *arXiv:1004.2515*, April 2010.

[18] Christopher J. Ellison, John R. Mahoney, and James P. Crutchfield. Prediction, Retrodiction, and The Amount of Information Stored in the Present. *Journal of Statistical Physics*, 136(6):1005–1034, 2009.

[19] Robin C. Ball, Marina Diakonova, and Robert S. MacKay. Quantifying Emergence in Terms of Persistent Mutual Information. *Advances in Complex Systems*, 13(3):327–338, 2010.

[20] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.