Multiscale Structure of More-than-Binary Variables

Blake C. Stacey[1]

[1] *Physics Department, University of Massachusetts Boston*

(Dated: May 10, 2017)

In earlier work, my colleagues and I developed a formalism for using information theory to understand scales of organization and structure in multi-component systems. One prominent theme of that work was that the structure of a system cannot always be decomposed into pairwise relationships. In this brief communication, I refine that formalism to address recent examples which bring out that theme in a novel and subtle way. After summarizing key points of earlier papers, I introduce the crucial new concept of an *ancilla component,* and I apply this refinement of our formalism to illustrative examples. The goals of this brief communication are, first, to show how a simple scheme for constructing ancillae can be useful in bringing out subtleties of structure, and second, to compare this scheme with another recent proposal in the same genre.

Imagine a system that is made up of a large number of pieces. A theme one encounters in many areas of science is that such a system is simpler to understand when the pieces are independent of one another, each one "doing its own thing." In thermodynamics and statistical physics, for example, the study of fluids begins with the conceptual model known as the ideal gas, in which the atoms sail past each other without interacting. On the other hand, a system can also be simple to understand if all the pieces are so tightly correlated or so strongly bound together that they essentially move as a unit. Then, the task of understanding the system again simplifies, because knowing what any one part is doing tells us most of what we need to know about the whole.

The general approach that my colleagues and I have developed over recent publications is that we can use *information theory* to make these heuristic discussions quantitative [1–3]. One conclusion is that we should not aim to quantify a system's "complicatedness" by a single number. Instead, it is more illuminating to devise *curves* that indicate how much structure is present *at different scales.* Another general theme is that one can mathematically formalize the notion of describing a system by augmenting that system with extra components and applying information theory to the augmented system. In this note, I will review the relevant aspects of information theory and the formal concept of system description, and then I will refine our earlier treatment of system description [1] to make it sensitive to complex structure in new ways.

We start by saying that a system $\mathcal{S}$ is composed of pieces, or components. We will denote the set of components of the system $\mathcal{S}$ by S. It is important to distinguish the two, because we could take the same components and arrange them in a different way to create a different system. We express the arrangement or the patterning of the components by defining an **information function**. For each subset $T \subset S$, the information function H assigns a nonnegative number $H(T)$, which expresses how much information is necessary to describe exactly the configuration or the behavior of all the components in the set T. One can prove a significant amount from the starting point that if H is to be an information function, it must satisfy a few basic axioms.

First, as we stated above,

$$H(T) \geq 0 \text{ for any } T \subset S. \tag{1}$$

Second,

$$\text{if } T \subset V \subset S, \text{ then } H(T) \leq H(V). \tag{2}$$

We call this property **monotonicity**.

Third, if we have two subsets $T \subset S$ and $V \subset S$, the total information assigned to their union, $H(T \cup V)$, must be limited. Information pertinent to the components in T can be pertinent to the components in V. For one reason, the sets T and V might have some components in common. And even those components that are not shared between the two sets might be correlated in some way such that the amount of information necessary to describe the whole collection is reduced. So, we require that

$$H(T \cup V) \leq H(T) + H(V) - H(T \cap V), \tag{3}$$

which we call **strong subadditivity**. Given two components s and t within S, the **shared information** expresses the difference between what is required to describe them separately versus describing them together:

$$I(s;t) = H(\{s\}) + H(\{t\}) - H(\{s,t\}). \tag{4}$$

It follows from strong subadditivity that the shared information is always nonnegative.

A **descriptor** of a system is an entity outside that system which tells us about it in some way. Mathematically speaking, we take the system $\mathcal{S}$ and augment it with a new component that we can call d, to form a new system whose component set is $S \cup \{d\}$. The information function of the augmented system reduces to that of the original system $\mathcal{S}$ when applied to subsets of S. The values of the augmented system's information function on subsets that include the descriptor d tell us how d shares information with the original components of $\mathcal{S}$.

The **utility** of a descriptor d is

$$U(d) = \sum_{s \in S} I(d;s). \tag{5}$$

Given the basic axioms of information functions that we listed above, we can define an **optimal descriptor** as the one which has the largest possible utility, given its own amount of information. That is, if we invest an amount of information y in describing the system $\mathcal{S}$, then an optimal descriptor has $H(d) = y$, and it relates to $\mathcal{S}$ in such a way that $U(d)$ is as large as the basic axioms of information functions allow. This defines a linear programming problem whose solution is the **optimal utility**, and so the theory of linear programming lets us prove helpful results about how the optimal utility varies as a function of y. Taking the derivative of the optimal utility gives the **marginal utility of information**, or MUI.

We proved several useful properties of the MUI in an earlier article [1]. For systems of the ideal-gas type, where information applies to one component at a time, the MUI is a low and flat function: Investing one bit of description buys one unit of utility, until the whole system is described. On the other hand, if all the components are bound together and "move as one," then investing a small amount of information buys us utility on a large scale, because that small amount applies across the board. In this case, the MUI starts high and falls sharply.

The MUI is defined using a single descriptor component. It is natural to speculate that constructions involving multiple descriptors could provide useful elaborations of the multiscale complexity formalism. This is one motivation for the developments that follow.

When the construction of a system is specified in detail, it is sometimes possible to make a finer degree of analysis, which reveals features that a first application of a structure index can overlook. To illustrate this, consider a system defined by a set of random variables, to which we ascribe some joint probability distribution. When systems are defined in this way, we can use the **Shannon information** (a.k.a., Shannon entropy, Shannon index) as our information function H. In the absence of an external reference to compare the values of these variables against, the most obvious meaningful statement we can make about them is whether the values are equal. (If the numbers recorded masses in grams, for example, then the difference between "0" and "3" would be more dramatic than that between "0" and "1," but we do not know that *a priori*. Information theory has, in certain ways, neglected the idea that some differences between symbols are more striking than others [4, 5].) Let us say that the system has three components, a_1, a_2 and a_3. Following the general idea of adding a descriptor to the system, as we did with the MUI, we introduce a new variable Δ_{12} which takes the value 1 when the state of a_1 and a_2 are the same, and is 0 otherwise. This new **ancilla** variable is determined completely by the original system, and is sensitive to the particular values taken by the original system components a_1, a_2 and a_3. We can define two other ancillae in the same way, Δ_{13} and Δ_{23}. Then, we can use the tools of multiscale information theory, like the MUI, to explore the structure of the ancilla variables, which in turn tells us about the structure of the original system. (The term "ancilla" is common in *quantum* information theory, in a sense similar to this [6–8].)

As a preliminary, let us try this with a three-component **parity-bit system**, which we can think of as picking a row at random from the XOR truth table:

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}. \tag{6}$$

Then, the values of the ancillary variables Δ_{12}, Δ_{13} and Δ_{23} in each possible joint state are the rows in the matrix

$$\Delta_X = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{7}$$

This is, again, a parity-bit system, but with odd parity instead of even. We can also think of it as the truth table of the XNOR logic gate: Any column is the NOT of the XOR of the other two. The fact that the structure does not simplify when we consider the variables pair-by-pair indicates that this system is not structured in a pairwise way, confirming what we noted in [1].

For a more elaborate example, take the two systems defined by James and Crutchfield [9]. These are three-component systems composed of random variables whose joint states are chosen by picking a row at random from a table, with uniform probability. The **dyadic** and

triadic systems are defined respectively by the tables

$$D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 3 \\ 2 & 1 & 0 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \\ 3 & 3 & 3 \end{pmatrix} \; ; \; T = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 2 & 2 \\ 1 & 3 & 3 \\ 2 & 0 & 2 \\ 3 & 1 & 3 \\ 2 & 2 & 0 \\ 3 & 3 & 1 \end{pmatrix} . \tag{8}$$

If we compute the MUI for these two systems in the manner described above, we find that the MUI for the dyadic system is the same as for the triadic. In fact, the information functions for the two systems, computed according to the Shannon scheme, agree for all subsets U. However, we can detect a difference between their structures by *augmenting* them, in the manner described above.

Specifically, if we take the dyadic example system and introduce three ancillae Δ_{12}, Δ_{13} and Δ_{23} in the manner described above, we find that the three ancillae form a completely correlated block system (that is biased towards the joint state 000). In contrast, for the triadic example, defining three ancillae in the same way, we find that they form a parity-bit system (with odd parity). This reveals that the tradic system has an information-theoretic structure at the scale of three variables in a way that the dyadic system does not. Explicitly,

$$\Delta_D = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \; ; \; \Delta_T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} . \tag{9}$$

The ancillary system defined by Δ_T has four possible distinct joint states, all of which are equally probable, and so it has two bits of information overall. For each possible joint state of the ancillary system, the original system can be in one of two joint states, with equal probability. Therefore, we see that the three bits of information necessary to specify the state of the original system break down into a pair of bits for describing the ancillae, plus one more bit of additional detail. For both the dyadic and triadic examples, the MUI of the ancillary systems Δ_D and Δ_T takes a simple form; in fact, these cases were both solved in [1]. For the ancillae of the dyadic system, Δ_D, if we define

$$h = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) \approx 0.8113, \tag{10}$$

we have

$$M(y) = \begin{cases} 3, & y \leq h; \\ 0, & y > h. \end{cases} \tag{11}$$

And for the ancillae Δ_T of the triadic system, we have

$$M(y) = \begin{cases} \frac{3}{2}, & y \leq 2; \\ 0, & y > 2. \end{cases} \tag{12}$$

Since I first posted a note on this topic [10], Ince published an article on the arXiv that also addressed the scenarios posed by James and Crutchfield [11]. Ince's article is a good occasion to compare our approach to multiscale information (and the elaboration of it we have begun here) to other programs of research. In what follows, I will attempt to draw attention to certain distinctions, which I believe are better thought of as complementary, rather than contradictory, conceptual developments.

Another way we could have elaborated upon our earlier multiscale information formalism would be to define an ancilla for each possible value of each component of a system. We could designate this **exploding** the original system. Applying our theory to an exploded system can reveal new details of organization, at the price of increasing the number of components one must consider. Suppose again that we have N random variables, each one of which has K internal states. If we explode the system and invent an indicator variable for each possible internal state of each original variable, then we have NK indicator variables, and specifying an information function on that set requires fixing $2^{NK} - 1$ numbers.

For example, consider a two-component system defined by picking two successive characters at random out of a large corpus of English text. The shared information between the two components quantifies how much knowing the value of the first character helps us predict the value of the second. However, knowing that the first character is a Q is a stronger constraint than knowing that it is, say, a T, because fewer characters can follow a Q. We can express this in our theory by exploding the two-component system, defining new components that represent the events of the first character being a Q and the first character being a T. This is the same basic idea as that of the **partial information decomposition** [12]. The raw data from which the PID is constructed is a set of values of "specific information," which is a measure of the information that a source variable provides about a particular outcome of a target variable. The specific information between a source X and a specific value y of a target Y, denoted $I(Y = y; X)$, measures the amount by which learning the value of X makes observing the value y become less surprising. By construction, averaging $I(Y = y; X)$ over all values of the variable Y, weighted by their probabilities, yields the standard mutual information $I(Y; X)$. The PID requires the user to introduce a distinction between "input" and "output" variables. Of course, in some scenarios, such a distinction is a natural one to introduce, but (as James and Crutchfield note) that is not true everywhere.

Ince proposes a **partial entropy decomposition**, inspired by the PID but without the distinction between "input" and "output" variables. Like our treatment with ancilla variables, Ince's PED can identify the parity-bit structure within the triadic example system.

The PED comes in two versions, both of which have some features that strike me as potential loci for future invention. In one version of the PED, the meaning of the terms in the expansion can be ambiguous; in the other, a desirable property of the original PID is lost. A further issue, common to both versions, is that one step in the definition explicitly requires the use of probabilities, and so it can only be formulated for the Shannon index. As we have argued elsewhere [1], the concept of "information" ought to be considered more generally. Rather than founding everything on logarithms of probabilities, it is beneficial to prove as much as possible starting from *the properties that a reasonable measure of information ought to satisfy.* (This approach is perhaps closest in spirit to that of Quax, Har-Shemesh and Sloot [13], which likewise appeared in a journal after the first version of this note was posted [10].) It would be interesting to see how the PED might be formulated in this manner.

Finally, note that Ince develops the PED for three-component systems, and demonstrates its usefulness there. As Ince points out, extensions to larger systems require making some

decisions on a conceptual level, but they should be quite interesting as well. I suspect that a PED that is viable for larger systems could open the way to merging the concepts of the PID/PED and the MUI.

My colleagues have been advocating the study of multiscale structure for a long time [14–16], and so I am grateful to all those who are raising new challenges in the subject, thereby demonstrating its intellectual vitality.

---

[1] B. Allen, B. C. Stacey and Y. Bar-Yam, "An Information-Theoretic Formalism for Multiscale Structure in Complex Systems," arXiv:1409.4708 [cond-mat.stat-mech] (2014).

[2] B. C. Stacey, *Multiscale Structure in Eco-Evolutionary Dynamics*. PhD thesis, Brandeis University (2015). arXiv:1509.02958 [q-bio.PE].

[3] B. C. Stacey, B. Allen and Y. Bar-Yam, "Multiscale information theory for complex systems: Theory and applications." In *Information and Complexity*, M. Burgin and C. S. Calude, eds. (World Scientific, 2017.)

[4] B. Allen, M. Kon and Y. Bar-Yam, "A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats," *The American Naturalist* **174,** 2 (2009), 236–43. http://www.necsi.edu/research/evoeco/.

[5] T. Leinster and C. A. Cobbold, "Measuring diversity: the importance of species similarity," *Ecology* **93,** 3 (2012), 477–89. http://www.maths.gla.ac.uk/~cc/pdf/Leinster2011.pdf.

[6] A. Peres, "Neumark's theorem and quantum inseparability," *Foundations of Physics* **20,** 12 (1990), 1441–53.

[7] C. M. Caves, C. A. Fuchs and R. Schack, "Unknown Quantum States: The Quantum de Finetti Representation," *Journal of Mathematical Physics* **43,** 9 (2002), 4537–59, arXiv:quant-ph/0104088.

[8] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, 2011).

[9] R. G. James and J. P. Crutchfield, "Multivariate dependence beyond Shannon information," arXiv:1609.01233 [cs.IT] (2016).

[10] B. C. Stacey, "Multiscale structure, information theory, explosions," https://www.sunclipse.org/?p=2257 (2017).

[11] R. A. A. Ince, "The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal," arXiv:1702.01591 [cs.IT] (2017).

[12] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," arXiv:1004.2515 [cs.IT] (2010).

[13] R. Quax, O. Har-Shemesh and P. M. A. Sloot, "Quantifying Synergistic Information Using Intermediate Stochastic Variables," *Entropy* **19,** 2 (2017), 85.

[14] Y. Bar-Yam, *Dynamics of Complex Systems* (Addison-Wesley, 1997).

[15] Y. Bar-Yam, "Multiscale complexity/entropy," *Advances in Complex Systems* **7** (2004), 47–63. http://necsi.edu/research/multiscale/.

[16] S. Gheorghiu-Svirschevski and Y. Bar-Yam, "Multiscale analysis of information correlations in an infinite-range, ferromagnetic Ising system," *Physical Review E* **70,** 6 (2004), 066115.