

## Network Information and Connected Correlations

Elad Schneidman,<sup>1,2,3</sup> Susanne Still,<sup>1,3</sup> Michael J. Berry II,<sup>2</sup> and William Bialek<sup>1,3</sup>

<sup>1</sup>*Department of Physics, Princeton University, Princeton, New Jersey 08544, USA*

<sup>2</sup>*Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA*

<sup>3</sup>*Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA*

(Received 15 July 2003; published 2 December 2003)

Entropy and information provide natural measures of correlation among elements in a network. We construct here the information theoretic analog of connected correlation functions: irreducible  $N$ -point correlation is measured by a decrease in entropy for the joint distribution of  $N$  variables relative to the maximum entropy allowed by all the observed  $N - 1$  variable distributions. We calculate the “connected information” terms for several examples and show that it also enables the decomposition of the information that is carried by a population of elements about an outside source.

DOI: 10.1103/PhysRevLett.91.238701

PACS numbers: 89.75.Hc, 05.50.+q, 05.70.Ce, 89.70.+c

In statistical physics and field theory, the nature of order in a system is characterized by correlation functions. These ideas are especially powerful because there is a direct relation between the correlation functions and experimental observables such as scattering cross sections and susceptibilities. As we move toward the analysis of more complex systems, such as the interactions among genes or neurons in a network, it is not obvious how to construct correlation functions which capture the underlying order. On the other hand it is possible to observe directly the activity of many single neurons in a network or the expression levels of many genes, and hence real experiments in these systems are more like Monte Carlo simulations, sampling the distribution of network states.

Shannon proved that, given a probability distribution over a set of variables, entropy is the unique measure of what can be learned by observing these variables, given certain simple and plausible criteria (continuity, monotonicity, and additivity) [1]. By the same arguments, mutual information arises as the unique measure of the interdependence of two variables or two sets of variables. Defining information theoretic analogs of higher order correlations has proved to be more difficult [2–10]. When we compute  $N$ -point correlation functions in statistical physics and field theory, we are careful to isolate the connected correlations, which are the components of the  $N$ -point correlation that cannot be factored into correlations among groups of fewer than  $N$  observables. We propose here an analogous measure of connected information which generalizes precisely our intuition about connectedness and interactions from field theory; a closely related discussion for quantum information has been given recently [11].

Consider  $N$  variables  $\{x_i\}$ ,  $i = 1, 2, \dots, N$ , drawn from the joint probability distribution  $P(\{x_i\})$ ; this has an entropy [12],

$$S(\{x_i\}) = -\sum_{\{x_i\}} P(\{x_i\}) \log P(\{x_i\}). \quad (1)$$

The fact that  $N$  variables are correlated means that the entropy  $S(\{x_i\})$  is smaller than the sum of the entropies for each variable individually,

$$S(\{x_i\}) < \sum_i S(x_i). \quad (2)$$

The total difference in entropy between the interacting variables and the variables taken independently can be written as [2,3]

$$I(\{x_i\}) \equiv \sum_i S(x_i) - S(\{x_i\}) = \sum_{\{x_i\}} P(\{x_i\}) \log \left[ \frac{P(\{x_i\})}{\prod_j P_j(x_j)} \right], \quad (3)$$

which is the Kullback-Leibler divergence between the true distribution  $P(\{x_i\})$  and the “independent” model formed by taking the product of the marginals,  $\prod_j P_j(x_j)$ . This has been called the multi-information; it provides a general measure of nonindependence among multiple variables in a network.

The multi-information alone does not tell us how much of the nonindependence among  $N$  variables is intrinsic to the full  $N$  variables and how much can be explained from pairwise, triple, and higher order interactions. For example, if the  $x_i$ 's are binary variables or equivalently Ising spins  $\sigma_i$ , and if the full distribution  $P(\{\sigma_i\})$  is a conventional Ising model with pairwise exchange interactions, then in an obvious sense there is nothing “new” to learn by observing triplets of spins that cannot be learned by looking at all the pairs. On the other hand, if  $\sigma_3$  is formed as the exclusive OR (XOR) of the variables  $\sigma_1$  and  $\sigma_2$ , then the essential structure of  $P(\sigma_1, \sigma_2, \sigma_3)$  is contained in a three-spin interaction; if  $\sigma_1$  and  $\sigma_2$  are chosen at random as inputs to the XOR, then all pairwise mutual informations among the  $\sigma_i$  will be zero, although the multi-information will be 1 bit (Fig. 1).

What we would like to do in our example of three variables is to separate that component of  $I(x_1; x_2; x_3)$  which is expected from observations on pairs of variables

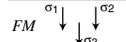
	$I(\sigma_1; \sigma_2; \sigma_3)$	$I_C^{(3)}$	$I_C^{(2)}$	$I(\sigma_1; \sigma_2)$	$I(\sigma_1; \sigma_3)$	$I(\sigma_2; \sigma_3)$	$R$ (or $I_3$ )
	0.8113	0	0.8113	0	0.3113	0.3113	-0.1887
	0.8113	0	0.8113	0	0.3113	0.3113	-0.1887
	1	1	0	0	0	0	-1
	2	0	2	1	1	1	1

FIG. 1. The values of multi-information, connected information of orders 2 and 3, the pairwise mutual information, and pairwise redundancy for three binary variables, whose probability distribution is given by the logical functions AND, OR, and XOR (with the inputs  $\sigma_1$  and  $\sigma_2$  chosen at random), and the case of ferromagnetic interaction, FM.

from that component which is intrinsic to the triplet. Observing the variables in pairs means that we can construct all of the pairwise marginals  $P_{ij} = \sum_{x_k} P(x_i, x_j, x_k)$ . Knowledge of these marginals provides (in general) a partial characterization of the full probability distribution  $P(x_1, x_2, x_3)$ . Following Jaynes [13] we can quantify this knowledge by saying that the pairwise marginals set a maximum value of the entropy for the full distribution. More generally, if we have  $N$  variables and we observe all the subsets of  $k$  elements, then there is a maximum entropy for the distribution  $P(\{x_i\})$  that is consistent with all of the  $k$ th order marginals. Let us write this maximum entropy distribution by  $\tilde{P}^{(k)}(\{x_i\})$  and denote the entropy of a probability distribution by  $S[P]$ ; note that

$$\tilde{P}^{(1)}(\{x_i\}) = \prod_{i=1}^n P_i(x_i), \quad (4)$$

and that  $\tilde{P}^{(N)}(\{x_i\})$  is just the true distribution  $P(\{x_i\})$ . Then we can decompose the multi-information among the  $N$  variables into a sequence of terms:

$$I(\{x_i\}) \equiv S\left[\prod_{i=1}^N P_i(x_i)\right] - S[P(\{x_i\})] = \sum_{k=2}^N I_C^{(k)}(\{x_i\}), \quad (5)$$

where we define the connected information of order  $k$ ,

$$I_C^{(k)}(\{x_i\}) = S[\tilde{P}^{(k-1)}(\{x_i\})] - S[\tilde{P}^{(k)}(\{x_i\})]. \quad (6)$$

The connected information of order  $k$  is positive or zero; it represents the amount by which the maximum possible entropy of the system decreases when we go from knowing only the marginals of order  $k-1$  to knowing also the marginals of order  $k$ . Each time that we increase the number of elements that we can observe simultaneously we uncover a potentially richer set of correlations, leading to a reduction in the maximum possible entropy; the connected information measures this entropy reduction.

Computing the connected information requires that we construct the maximum entropy distributions consistent with marginals of order  $k$ . In general this is a difficult problem. Recall that to maximize the entropy when we know the expectation values of functions  $F_\mu(\{x_i\})$ , the

resulting probability distribution is of the Boltzmann form,  $P(\{x_i\}) \propto \exp[-\sum_\mu \lambda_\mu F_\mu(\{x_i\})]$ , where the  $\lambda_\mu$  are Lagrange multipliers conjugate to each function [13]. We can think of each marginal distribution as a set of expectation values over the full distribution, so that we need one Lagrange multiplier for each  $k$ -tuple of  $x$  values. The distribution  $\tilde{P}^{(k)}$  thus has the form of a Boltzmann distribution with  $k$ -body interactions; these interactions are arbitrary functions which have to be determined by matching the observed marginals. As an example, for three variables with known pairwise marginals the maximum entropy distribution takes the form

$$\tilde{P}^{(2)}(x_1, x_2, x_3) = \frac{1}{Z} \exp[-\lambda_{12}(x_1, x_2) - \lambda_{23}(x_2, x_3) - \lambda_{31}(x_3, x_1)]. \quad (7)$$

For a physical system that has at most  $K$ -body interactions among the  $N$  variables,  $\tilde{P}^{(K)}$  will be the exact distribution. Correspondingly,  $I_C^{(k)} = 0$  for  $k > K$ .

In general the functions  $\lambda$  are difficult to determine from the observed marginals, but this is not the case for  $k=1$ . This is a well known but important point: the maximum entropy distribution consistent with one-body marginals is just the product of the marginals, but the maximum entropy distribution consistent even with two-body (pairwise) marginals is not simply written in terms of the marginals because the observed two-body correlations include an average over interactions with all other degrees of freedom. As a result, even the second order maximum entropy distributions for  $N$  variables are not simply related to the pairwise marginals, and the second order connected information is not simply related to the mutual information among pairs of variables;  $I_C^{(2)}$  is larger than the mutual information between any pair of variables, but is not equal to their sum.

The fact that maximum entropy distributions have an exponential form, and in the binary or Ising case this form includes only a finite set of parameters, connects our discussion with previous work. A number of authors have used the maximum entropy distribution for families of parametrized models as part of statistical tests for the existence of higher order interactions [4,8,14]. In related work, Amari [9] has constructed a geometry on the parameter space for exponential families using the Fisher information as a metric, and in this geometry the maximum entropy distributions are orthogonal projections onto subspaces of the full parametric space (see also [5]). Rather than providing a parametric model of  $k$ th order interactions and determining a confidence level, the set of  $I_C^{(k)}$  provides a quantitative characterization of the relative importance of various order interactions, independent of parametrization.

As examples (Fig. 1), consider three binary or Ising variables related either by Boolean functions (AND, OR, XOR) or coupled through a pairwise ferromagnetic interaction (FM). For these simple functions, we find that the

multi-information is composed of either pure two-body interactions or pure three-body ones, as our intuition suggests. When we add noise either to the input or output of the Boolean functions (Fig. 2) we degrade the correlations, but more interestingly we find that pure two-body interactions such as AND and OR show a three-body interaction component for some types of noise (even for noise sources which are state dependent). For the pure three-body XOR, noise may result in the appearance of two-body interactions. For these three functions, input noise changes only the strength of the existing interactions, rather than introducing a new kind of effective interaction.

As is familiar in physical examples, if we observe only some of the elements of a network then the effect of the hidden elements may be to create new effective interactions among the observed elements. As examples (Fig. 3), when one hidden binary element determines the nature of pure pairwise interaction among the remaining elements, the observable subnetwork can have an effective three-body interaction. Alternatively, for a network with only pure three-body interactions, hidden elements can induce an effective two-body interaction among the observables.

As noted above, the connected information at second order (for example) cannot be written simply in terms of the mutual information among pairs of variables. Many previous authors have looked for linear combinations of mutual information measures which might provide mea-

sures of higher order interaction, and among these one approach is of particular interest [2,3,6,7,10]: If we draw a Venn diagram of regions in the plane corresponding to the variables  $x_1, x_2, x_3$ , identifying areas with entropies, then the mutual information  $I(x_i; x_j)$  between two variables is the area of their intersection, and there is a unique region shared by all three variables; with the area-entropy correspondence the size of this “triplet information” is

$$I_3 = \sum_i S(x_i) - \sum_{i<j} S(x_i, x_j) + S(x_1, x_2, x_3) = I(x_1; x_2; x_3) - \sum_{i<j} I(x_i; x_j). \quad (8)$$

This proposal for measuring a pure triplet information has natural generalizations to more than three variables.

There are at least two difficulties with the triplet information defined by  $I_3$  (see a thorough discussion in [3]). First, despite the identification of shared information with areas in the plane, we find that  $I_3$  can be negative (AND, OR, and XOR in Fig. 1). Second,  $I_3$  can be nonzero even for networks that have only pairwise interactions (FM in Fig. 1).

Rather than “triplet information,”  $I_3$  actually compares [2,3] the information that  $x_1$  and  $x_2$  together provide about  $x_3$  with the information that these two variables provide separately:

$$I_3 = [I(x_1; x_3) + I(x_2; x_3)] - I(\{x_1, x_2\}; x_3). \quad (9)$$

This comparative measure of information is symmetric

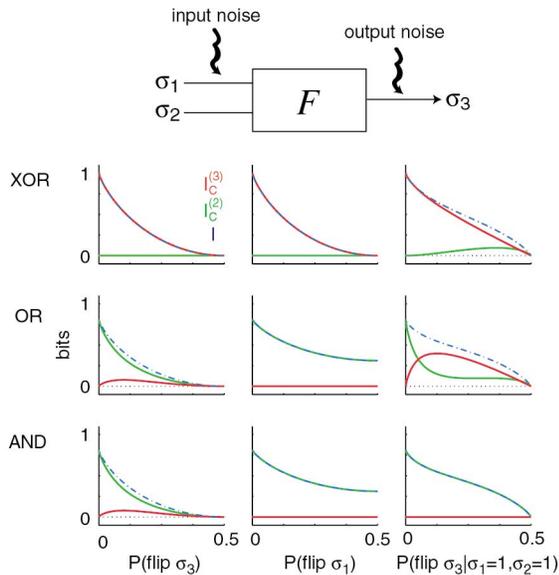


FIG. 2 (color). Correlated information of orders 2 and 3 and the multi-information for three variables whose joint probability distribution is given by noisy logical functions. Each panel presents the  $I_C$ 's and  $I$  values for a noisy version of one Boolean gate (XOR in first row, OR in second, AND in third), as a function of noise amplitude. The three types of noise are output noise (probability of flipping  $\sigma_3$ ), input noise (probability of flipping  $\sigma_1$ ), and input-dependent output noise (probability of flipping  $\sigma_3$ , given that  $\sigma_1 = 1$  and  $\sigma_2 = 1$ ).

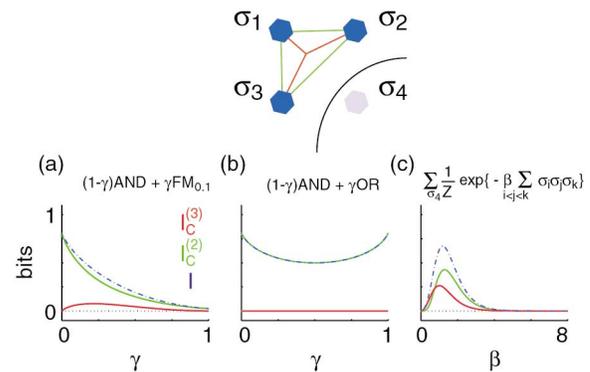


FIG. 3 (color). Correlated information of orders 2 and 3 and the multi-information, for networks of three binary observable elements,  $\sigma_1, \sigma_2, \sigma_3$ , with one hidden binary element  $\sigma_4$ . (a)  $I_C$ 's and  $I$  values for a network where the value of  $\sigma_4$  determines the pairwise interaction between the other elements: if  $\sigma_4 = 0$  then  $\sigma_3 = \text{AND}(\sigma_1, \sigma_2)$ ; if  $\sigma_4 = 1$  then the interaction among the observable variables is (pairwise) ferromagnetic with a finite temperature ( $\beta = 0.1$ ). Information values are plotted as a function of  $\gamma = P(\sigma_4 = 0)$ .  $P(\sigma_1, \sigma_2, \sigma_3)$  in (b) same as (a), but for a  $\sigma_4$ -dependent mixture of AND and OR. For this case there is no effective three-body interaction. (c)  $I_C$ 's and  $I$  for the three observable binary variables network, where the full four element network has pure three-body interactions, plotted as a function of the inverse temperature  $\beta$ .

under permutation of the indices, so the labeling of variables as 1, 2, 3 is arbitrary. If  $I_3$  is positive, then any pair  $x_i$  and  $x_j$  are redundant in terms of the information that they provide about the remaining  $x_k$ . If  $I_3$  is negative, then there is synergy—two variables taken together are more informative than they are when taken separately.

The question of synergy and redundancy brings us back to one of the primary motivations for this analysis. Consider the responses  $x_1, x_2, \dots, x_N$  of a collection of elements to some stimulus  $y$ —for example, a group of neurons responding to a sensory stimulus. For each neuron  $i$  we can ask how much information the response provides about the sensory world,  $I(x_i; y)$ . When we look at a pair of neurons, we can ask whether these neurons provide redundant or synergistic information [using Eq. (9); see, e.g., [15,16]]. Similarly for a large population of neurons we can compare the information in the population,  $I(\{x_i\}; y)$ , with the sum of informations provided by the neurons individually,  $\sum_i I(x_i; y)$ . This comparison, however, does not tell us whether (for example) the synergy in the population is the result of pairwise correlations or whether there are special combinations of responses across all three or more neurons which provide extra information. The possible significance of such multineuron combinatorial events has been discussed for many years (see, e.g., [17–19]).

We recall that the information provided by a population of neurons can be written as

$$I(\{x_i\}; y) = S[P(\{x_i\})] - \langle S[P(\{x_i\}|y)] \rangle_y, \quad (10)$$

where  $\langle \dots \rangle_y$  denotes an average over the distribution of sensory inputs. The redundancy of the population is defined as

$$R(\{x_i\}) \equiv \sum_{i=1}^N I(x_i; y) - I(\{x_i\}; y), \quad (11)$$

where negative  $R$  corresponds to synergy. We note that  $R$  can be written as the difference between two multi-information terms,

$$R(\{x_i\}) = \left( \sum_{i=1}^N S[P(x_i)] - S[P(\{x_i\})] \right) - \left\langle \sum_{i=1}^N S[P(x_i | y)] - S[P(\{x_i\} | y)] \right\rangle_y. \quad (12)$$

The first term is the multi-information in the distribution of neural responses, which measures the extent to which the total “vocabulary” of the population is reduced through correlations, while the second term is the multi-information in the distribution of responses to a given stimulus. Each of these terms in turn can be expanded as a sum of connected informations, so that

$$R(\{x_i\}) = \sum_{k=2}^N [I_C^{(k)}(\{x_i\}) - \langle I_C^{(k)}(\{x_i\}|y) \rangle_y], \quad (13)$$

where  $I_C^{(k)}(\{x_i\}|y)$  is the connected information of order  $k$  in the network of  $\{x_i\}$ , for a given value of  $y$ . By analogy with the discussion of synergy in pairs [16], the terms  $I_C^{(k)}(\{x_i\})$  quantify the contribution of  $k$ th order interactions to restricting the vocabulary of the population response (much as not all  $k$  letter combinations form words in English), while the terms  $\langle I_C^{(k)}(\{x_i\}|y) \rangle_y$  quantify the contribution of  $k$ th order correlations to reducing the noise in the population response.

To summarize, the maximum entropy construction of connected information presented here provides us both with a method for decomposing the correlations within a network and for quantifying the contribution of these correlations to the information that network states can provide about external signals. Since any part of a network can be thought of as “external” to its complement, this unified discussion of internal correlations and the representation of external signals is attractive.

This work was supported in part by the Pew Scholars Program and the E. Mathilda Ziegler Foundation (M. J. B.), the Rothschild Foundation (E. S.), and the German Research Foundation (S. S.). We thank A. Bell, I. Nemenman, and W. K. Ma.

- 
- [1] C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).
  - [2] W. J. McGill, *IRE Trans. Info. Theory* **4**, 93 (1954).
  - [3] S. Watanabe, *IBM J. Res. Dev.* **4**, 66 (1960).
  - [4] I. J. Good, *Ann. Math. Stat.* **34**, 911 (1963).
  - [5] I. Csizár, *Ann. Probab.* **3**, 146 (1975).
  - [6] T. S. Han, *Inform. Control* **36**, 113 (1978).
  - [7] M. Studený and J. Vejnarová, in *Learning in Graphical Models*, edited by M. I. Jordan (Kluwer, Dordrecht, 1988).
  - [8] L. Martignon, G. Deco, K. Laskey, M. Diamond, W. Freiwald, and E. Vaadia, *Neural Comput.* **12**, 2621 (2000).
  - [9] S. Amari, *IEEE Trans. Inf. Theory* **47**, 1701 (2001).
  - [10] A. Bell, Redwood Neuroscience Institute Technical Report No. 02-1, 2002.
  - [11] N. Linden, S. Popescu, and W. K. Wootters, *Phys. Rev. Lett.* **89**, 207901 (2002).
  - [12] We write a sum to indicate sums or integrals as appropriate.
  - [13] E. Jaynes, *Phys. Rev.* **106**, 620 (1957).
  - [14] A. Soofi, *J. Am. Stat. Assoc.* **87**, 812 (1992).
  - [15] T. Gawne and B. J. Richmond, *J. Neurophysiol.* **13**, 2758 (1993).
  - [16] N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. R. de Ruyter van Steveninck, *Neural Comput.* **12**, 1531 (2000).
  - [17] G. Palm, A. Aertsen, and G. Gerstein, *Biol. Cybern.* **59**, 1 (1988).
  - [18] M. Abeles, *Corticonics* (Cambridge University Press, Cambridge, 1991).
  - [19] M. J. Schnitzer and M. Meister, *Neuron* **37**, 499 (2003).