Accounting for synergy is essential for inferring information flow

Summary: Venkatesh et al. (2020) recently proposed a new framework for understanding flow of information about a "message" in a neural circuit, where the message could represent a stimulus or a response. Despite its novelty, their analysis was purely theoretical and did not address neuroscientific complexities in much detail. Through simulations on neuronal networks, we show that their theory stands the basic test of empirical validation and explain how their model can be employed in practice. We design networks of "theta" neurons to examine two common experimental situations: (1) an upstream brain region encodes information in the form of a spike-train, and is then processed by a downstream region, whose flows we wish to understand; (2) multiple populations encode and transmit information about a stimulus, and we sample only a subset of neurons in these populations. Our results, accompanied by models drawn from neuroscience, demonstrate that several aspects hypothesized by Venkatesh et al. can indeed be observed in realistic neural populations: (i) accounting for "synergy" is essential for seamlessly tracking information flow about a specific message; (ii) measuring edges rather than nodes can help avoid ambiguity in flow estimates; (iii) we can often use simplified correlation-based techniques to efficiently arrive at information-theoretically meaningful conclusions, while being aware of some caveats we describe.

Introduction: Recently, Venkatesh et al. (IEEE Trans. Inf. Thy. 2020) described a new framework for defining and inferring information flow about a specific message. They modeled the brain as a computational system within which they defined information flow tied to a "message" (e.g., a stimulus or a response), so as to capture the dynamics of how the flow may evolve. Their system was designed to allow for feedback and to give clearer interpretations in settings where Granger Causality provided ambiguous or incorrect answers. In proposing a new measure for information flow about a message, Venkatesh et al. make two important theoretical arguments: (i) it is necessary to account for "synergy" when detecting information flow; (ii) it is essential to measure edges to avoid ambiguities in flows. Here, "synergy" refers to the idea that two variables can *jointly* contain information about a message M, while *individually* containing *little or no* information about M (e.g., as measured by mutual information). This concept has been explored by many works in the literature in the context of *encoding* (e.g., see Schneidman et al., J Neurosci 2003; Gat & Tishby, NeurIPS 1999; Timme and Lapish, 2018). Venkatesh et al. (2020) formally connect synergy with *information flow* by developing an information-theoretic measure based on conditional mutual information, where *conditioning* plays the crucial role of *capturing synergy*. In the present work, we provide concrete simulations grounded in neuroscience to show how their information-theoretic measure may be estimated (approximately) in practice, and show how conditioning reveals synergy in experimental settings to support their main theoretical arguments.

Results: We present results from simulations of two neuronal network models, reflecting common experimental situations.

Model 1: We first consider a setting where a binary stimulus (or "message"), M, is encoded in the form of a spike train X_M by an upstream (e.g. sensory) region, and is transmitted for further processing to a downstream region. The downstream region is designed to show how synergy may arise in a neural circuit: it corrupts X_M with a noisy spike train X_Z, retains the noise X_Z along a separate path, and then recovers X_M by "subtracting" X_Z from the corrupted message (see Methods for details). Biological examples of such cancellation effects appear in the context of "corollary discharge" or "efference copies", for instance, feedback of eye-movement to vision, or the inability to tickle oneself (see Fukutomi & Carlson, *Front. Int. Neuro., 2020*, for a recent review). We are interested in tracking *where* information about M is present, at each time instant in the downstream network, and understanding which edges it flowed along.

Model 2: Next, we consider a setting where a binary stimulus M is encoded in the average firing rate of a population of neurons, P_M. After a delay, this population receives a corrupting noisy input P_Z from another population encoding a continuous noise variable Z. When there is a large amount of noise, P_Z is suppressed by a third inhibitory population after an axonal delay. We are interested in tracking the flow of M as it is processed in this circuit. We also consider subsamplings of these populations, where only a subset of the neurons in each population is recorded, as in a multi-electrode array recording.

In both aforementioned models, when information about M is encoded in the network in the form $X_M + X_Z$, *only upon conditioning* on X_Z are we able to statistically detect the presence of information flow in $X_M + X_Z$ (see p-values in Fig. 1c). We compute an approximation of Venkatesh et al.'s information flow measure at each node, within every 10ms time window. In Model 2, we use partial correlation, and in Model 1, we use mean absolute conditional correlation (MACC), as proxies for conditional mutual information. Partial correlation has a caveat: since it is the *average* of the conditional correlation over all values of the conditioning variables, i.e., since $\rho_{MX|Y} = \mathbb{E}_y[\rho_{MX|Y=y}]$, positive and negative conditional correlations may cancel. This is indeed what happens if we use partial correlation in Model 1, which is why we instead used the mean *absolute* conditional correlation, i.e., $\mathbb{E}_y|\rho_{MX|Y=y}|$. We compare these approximations of information flow to simple Pearson

correlation, and we find that in each Model, only the partial or conditional variant reveals where information flow is present at all time instants that the message is statistically discernible (see Fig. 1). This shows how accounting for synergy through conditioning is essential to detecting information flows in such circuits.

Another important theoretical argument propounded by Venkatesh et al. (2020) is that *edges* must be measured, i.e., we need to know both the *sending* node and the *receiving* node of each transmission. To see why, suppose two nodes X_1 and X_2 transmit the message M at time t=1, and two other nodes X_3 and X_4 transmit the same message at t=2. A node-only examination of flows, e.g. using Granger causality, would be unable to determine whether X_3 received the message from X_1 and X_4 from X_2, or X_3 from X_2 and X_4 from X_1, or some other combination of the two. Regression-based methods such as Granger causality implicitly overlook this issue, since they would assign a weight of half each to X_1 and X_2. Measuring and assigning information flow to *edges*, on the other hand, automatically resolves this ambiguity.

Methods: Simulations used reparametrized QIF ("theta") neuron models (Ermentrout and Kopell, SIAM 1986).

Model 1: The stimulus-dependent spike train X_M was generated using a random network of excitatory and inhibitory theta neurons, which was fed a constant current input based on M. The spike train itself was the time-dependent output of one neuron in this network which showed a sustained difference in firing rate for different values of M (Fig. 1b). The downstream network consists of three "nodes", each consisting of a group of neurons: X_1 and X_3 perform "XOR" (exclusive-or) operations, while node X_2 is a delay element. We simulate 1000 trials of spiking data, and compute p-values using a permutation test.

Model 2: The P_M population consisted of 200 excitatory and 200 inhibitory neurons with random interconnections. The P_Z population consisted of 100 excitatory and 100 inhibitory neurons, while the inhibitory population consisted of 100 excitatory neurons connected to 300 inhibitory neurons. The inhibitory population would activate and suppress P_Z (after a delay) if the firing in the P_M population exceeded a threshold. We simulated 100 trials of data and computed p-values using a standard chi-squared test.

XOR models such as in Model 1 have been considered in the past (e.g. Timme and Lapish, 2018; Gidon et al., Science 2020, give biological context), and are well known examples of synergy. However, Venkatesh et al. (2020) were the first to analyze synergy in the context of information *flow*, and our work is the first concrete demonstration of this. Although Timme and Lapish simulate and measure information quantities in simple XOR networks, they do not compute information flow across trials, as we do, and instead measure mutual information between two time windows, which is less clearly interpretable.

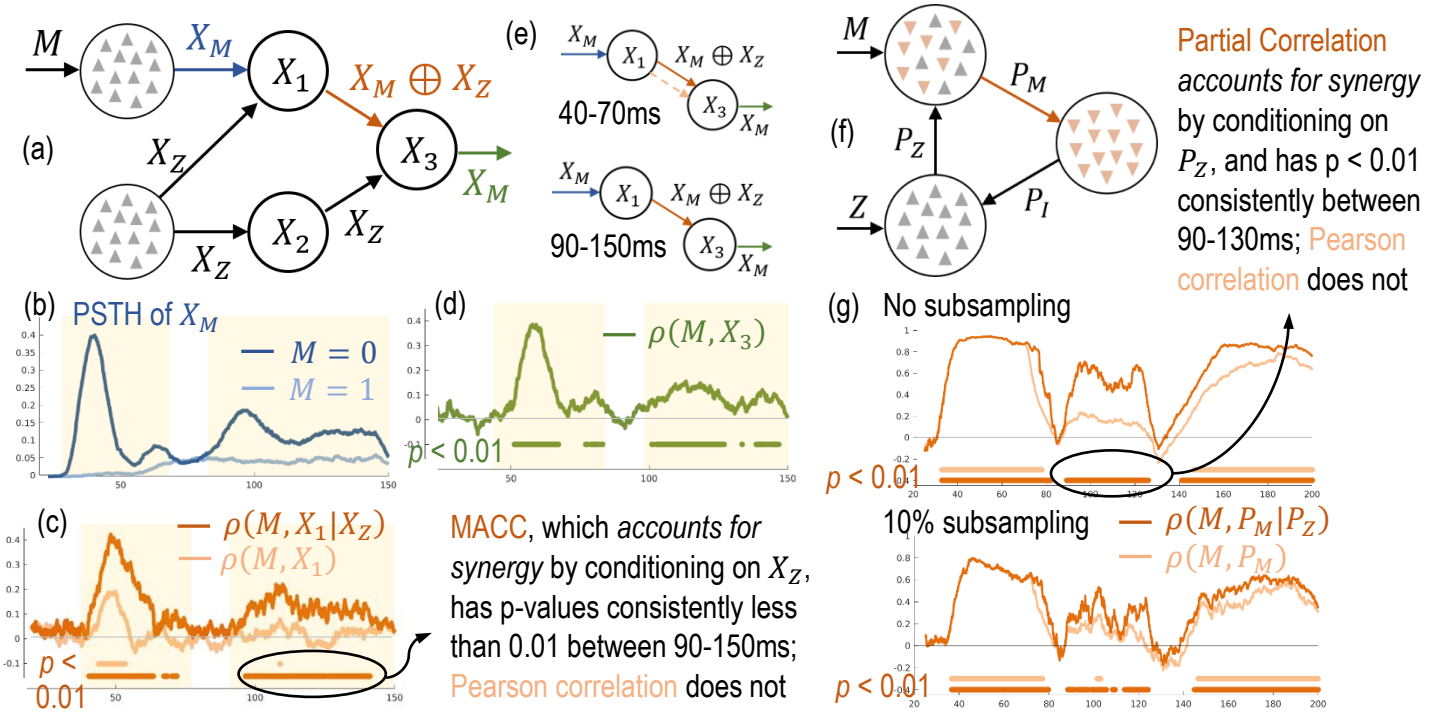

Figure 1: (a) A schematic of Model 1. (b) PSTH of X_M for M=0 and M=1. Highlighted regions show when M is discernible from X_M. (c,d) Correlation between M and transmissions of X_1 and X_3 respectively. Observe that between 90 and 150ms in (c), statistically significant information flow of M is seen only when using MACC, i.e., when conditioning on X_Z, as depicted in (e). (f) Schematic of the neural circuit in Model 2. (g) Correlation between M and transmissions shown in (f). Again, statistically significant information flow is seen between 100 and 150ms (at all levels of subsampling) only when using partial correlation.