Inference of Causal Information Flow in Collective Animal Behavior

Warren M. Lord, Jie Sun, Nicholas T. Ouellette, and Erik M. Bollt

***Abstract*—Understanding and even defining what constitutes animal interactions remains a challenging problem. Correlational tools may be inappropriate for detecting communication between a set of many agents exhibiting nonlinear behavior. A different approach is to define coordinated motions in terms of an information theoretic channel of direct causal information flow. In this work, we present an application of the optimal causation entropy (oCSE) principle to identify such channels between insects engaged in a type of collective motion called swarming. The oCSE algorithm infers channels of direct causal inference between insects from time series describing spatial movements. The time series are discovered by an experimental protocol of optical tracking. The collection of channels infered by oCSE describes a network of information flow within the swarm. We find that information channels with a long spatial range are more common than expected under the assumption that causal information flows should be spatially localized. The tools developed herein are general and applicable to the inference and study of intercommunication networks in a wide variety of natural settings.**

***Index Terms*—Bioinformatics, Biological systems, Inference algorithms, Graph theory, Nonlinear dynamical systems**

1. Introduction

COLLECTIVELY interacting groups of social animals such as herds, schools, flocks, or crowds go by many names depending on the specific animal species. But in all cases, they tend to display seemingly purposeful, coordinated group-level dynamics despite the apparent absence of leaders or directors. These coordinated group behaviors appear to emerge only from interactions between individuals, analogous to the manner in which macroscopic observables are determined by microscopic interactions in statistical physics. Thus, collective behavior has captivated a broad spectrum of researchers from many different disciplines [1]–[19].

Making the analogy to statistical physics more concrete, it is reasonable to suggest that a deep understanding of collective group motion may arise from three parallel pursuits. We can perform a macroscopic analysis, focusing on the observed group-level behavior such as the group morphology [20] or the material-like properties [18], [21], [22]; we can perform a microscopic analysis, determining the nature of the interactions between individuals [19], [23]–[25]; and we can study how the microscopic interactions scale up to give rise to the macroscopic properties [26].

The third of these goals—how the microscopic individual-to-individual interactions determine the macroscopic group behavior—has arguably received the most scientific attention to date, due to the availability of simple models of collective behavior that are easy to simulate on computers, such as the classic Reynolds [27], Vicsek [26], and Couzin [28] models. From these kinds of studies, a significant amount is known about the nature of the emergence of macroscopic patterns and ordering in active, collective systems [29]. But in arguing that such simple models accurately describe real animal behavior, one must implicitly make the assumption that the interactions between individuals are correctly represented. Any model of interactions has two key and distinct components: a specification of the mathematical form of the interaction, and, more fundamentally, a choice as to *which* individuals interact. Given that it is difficult to extract the appropriate social interaction network from empirical measurements, models typically replace this hard-to-measure social network with the simple-to-define proximity network [30]. Thus, it is assumed that individuals interact only with other animals that are spatially nearby. No matter what species is involved, the answer to the question of whether interactions are generally limited to or dominated by spatial local neighbors has strong implications. Recently, for example, scientists studying networks have shown that introducing even a small number of long range interactions into a lattice can impart qualitative changes to the observed macroscopic behavior [31], [32]. Consequently, the question of whether flocks or swarms or herds also contain long range interactions between individuals may have important implications for the understanding of collective motion.

Efforts to move past the simple framework of assuming that the local spatial neighborhood of an individual dominates its behavior have been largely theoretical [33], [34], as it is challenging to extract the underlying interaction network from measured data. Empirical methods have often relied upon various types of correlational (or other pairwise) time-series analysis [19], which by design only captures linear dependence and fail to detect the nonlinear relationships that are typical in real-world applications. An alternative paradigm would be to use information theoretic methods that are capable of detecting nonlinear dependence. One such example is Transfer Entropy [35], which is a type of conditional mutual information designed specifically for the detection of (possibly asymmetric) information flow between two coupled units. However, such methods also rely on *pairwise* computations, and thus cannot differentiate between direct and indirect interactions. As we

Warren M. Lord, Jie Sun, and Erik M. Bollt are with the Department of Mathematics, Clarkson University, Potsdam, New York 13699, USA. (email: lordwm@clarkson.edu, sunj@clarkson.edu, and bolltem@clarkson.edu)

Nicholas T. Ouellette is with the Department of Civil and Environmental Engineering, Stanford University, Stanford, California 94305, USA. (email: nto@stanford.edu)

have recently shown, any pairwise method, no matter how high its fidelity, will tend to overestimate the number of links in the interaction network, typically resulting in a significant number of false positives that cannot be resolved even with unlimited data [36], [37].

In this paper, we introduce a new mathematical framework based on optimal causation entropy (oCSE) to reveal the detailed network of interactions between individuals in a collective animal group. In brief, Causation Entropy (CSE) is capable of detecting and describing the interactions between three and higher numbers of components [36], [37]. That is, we can describe the influence of individual X on individual Y conditioned on the existence of the influence Z on Y. Thus, CSE allows us to draw conclusions quite different from what we could using pairwise interactions alone, in that information that "flows" from X to Y only through Z is an indirect influence that would at best be misclassified as a direct influence by a pairwise measure. Built upon the concept of CSE, oCSE is an efficient, constructive algorithm to infer the network of *direct* interactions based on CSE estimates [37]. In the application of inferring interactions among animals, oCSE requires knowledge only of the positions (or velocities or accelerations) of individuals in a group and is thus directly computable from empirical data. Because we define interactions via the information theoretic notion of the direct exchange of information as detected by uncertainty reduction, we need not make any assumptions about the spatial proximity of interacting individuals or the precise mathematical form of interaction. To demonstrate the unique utility of this oCSE network inference algorithm, we apply it to experimental measurements of the motion of individuals in mating swarms of the non-biting midge *Chironomus riparius*. In addition to showing the computability of the CSE in this data set, the oCSE approach clearly reveals that spatial proximity and interaction are not synonymous, suggesting that a deep understanding of collective behavior requires more subtle analysis of interactions than simple position-based proximity metrics.

2. Causation Entropy and the oCSE algorithm

We say roughly that information flows from agent X to agent Y if the future states of Y and the past states of X share information that cannot be accounted for by any other variables in the system. To make this statement precise we need to define the information associated with a variable, what it means to share information, and how to condition on other potential sources of information.

The information content descriptive of a random variable can be quantified by the Shannon entropy. Thus, if $p(x)$ is the probability that a measurement of a variable X take the particular value x, then the uncertainty associated with that insect's state is defined as its Shannon entropy [38],

$$H(X) \equiv -\sum_x p(x)\log p(x), \tag{2.1}$$

where the summation is taken over all possible values of X with the convention $p(x)\log p(x) = 0$ if $p(x) = 0$. The base of the logarithm is not important for inferring relationships between variables as long as the same base is adopted. If X is a real-valued continuous random variable, we may interpret $p(x)$ as the probability density function of X and use differential entropy, defined as $h(X) \equiv -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx$. In this section we use the notation H to mean (2.1); but $h(X)$ may also apply, depending on the application. Note that in the application to midge swarming, we assume that insect acceleration is a continuous variable.

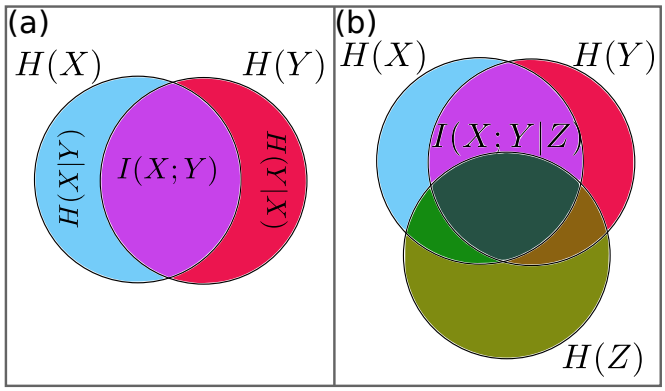

Fig. 1. Entropy and mutual information as visualized by Venn diagrams.

When two random variables are available the information in X can be subdivided into information belonging only to X and information belonging to both X and the other variable, Y. The mutual information describes the shared information between X and Y and is defined as [38], [39]

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{2.2}$$

Here $H(X,Y)$ denotes the entropy of the joint random variable (X,Y) whose measurements consist of ordered pairs, (x,y). Fig.(1a) shows a Venn diagram visualization of the relation between various entropies and mutual information.

Conditioning is similar to removing part of a circle in Fig.(1a) from the picture, leaving a crescent of the other circle remaining. For example, the conditional entropy of Y given X [38], [39],

$$H(Y|X) = H(X,Y) - H(X), \tag{2.3}$$

tells how much uncertainty is associated with Y given knowledge about X. The importance of conditional entropy for understanding swarm behavior is in finding the mutual information between two variables that is not present in a third variable. If Z is another discrete variable then the conditional mutual information of X and Y given Z is defined by

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z). \tag{2.4}$$

Fig.(1b) shows that conditional mutual information can be thought of as "removing" the part of information contributed by Z from the picture before computing the mutual information between X and Y.

Transfer Entropy is a type of conditional mutual information [35]. If information in X "flows" to Y then 1) there would have to be information contained in Y at a future time, say $t+\tau$, that is not explained by the state of Y at time t, and 2) this information would be shared by Y at time $t+\tau$ and X at time t. There are three variables present,

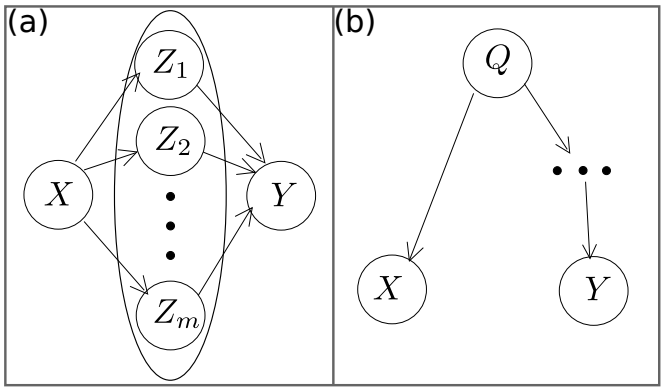

Fig. 2. Two ways that $I(X^{(t)}, Y^{(t+\tau)}|Y^{(t)})$ could be positive without an edge between X and Y in a discrete time dynamical process on a directed graph. (a) The variables Z_i (or even a subgraph of $\{Z_i\}_{i=1}^m$) serve as an intermediary between X and Y. (b) The states of X and Y are strongly influenced by a third source, Q. The "..." indicate that there might be other nodes on the path from Q to Y to induce the time lag τ.

which we can label $X^{(t)}$, $Y^{(t)}$, and $Y^{(t+\tau)}$. For information to "flow" from X to Y over time τ, then, it is necessary that $I(X^{(t)}; Y^{(t+\tau)}|Y^{(t)}) > 0$. The quantity on the left is proposed as Transfer Entropy $T_{X\to Y}$ because it reveals how much information has been transferred by the flow. This notion of an information flow, however, does not reflect how the information got there. When there are more than two insects present, information could flow from X to Y by going through a set of intermediaries [Fig.(2a)], or the information could have flowed to both from a third source, Q, but taken slightly longer to reach Y than X [Fig.(2b)].

Causation Entropy (CSE) is a tool designed to detect directed information flows. If X, Y and Z are variables related to three insects then the Causation Entropy of X to Y given Z is

$$C_{X\to Y|Z} = I(X^{(t)}; Y^{(t+\tau)}|Z^{(t)}). \tag{2.5}$$

In other words, $C_{X\to Y|Z}$ is the information shared between $X^{(t)}$ and $Y^{(t+\tau)}$ that is not already contained in $Z^{(t)}$. With the CSE so defined in terms of conditioning sets, then the question becomes how should the conditioning sets be chosen so as to reflect that *direct* information flow in the network. To this end, below we review oCSE as an algorithmic approach to efficiently learn the underlying interaction network structure.

The oCSE algorithm starts with an empty conditioning set and only adds as many variables as is necessary [37]. This is called the aggregation or discovery phase, which is followed by a removal phase in which redundancies are removed from the set [37]. To be more specific, let $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ and suppose that $\{Y\}\cup\mathcal{X}$ is a list of all variables of interest to the system. Initially, let $\mathcal{Z} = \varnothing$ be the set of potential causal variables. On each round the variable X_i is added to $\mathcal{Z}$ if

$$C_{X_i\to Y|\mathcal{Z}} = \max_{X_j\notin\mathcal{Z}} C_{X_j\to Y|\mathcal{Z}} > 0. \tag{2.6}$$

The discovery phase terminates when no such variable can be found from the remaining set of variables. The resulting set $\mathcal{Z}$ is possibly a superset of the variables that communicate directly with X because the value of $C_{X_i\to Y|\mathcal{Z}}$ can in fact be positive due to indirect information flow from X_i to Y, unless Z contains all the other true causal components [37]. Thus, the removal phase eliminates elements from $\mathcal{Z}$ if they are redundant given other elements in $\mathcal{Z}$. On each iteration a new member of $\mathcal{Z}$, Z_i, is chosen and removed if and only if

$$C_{Z_i\to Y|\mathcal{Z}\setminus\{Z_i\}} = 0. \tag{2.7}$$

In practice, when the CSEs are estimated from finite amount of data, the "$>$" and "$=$" in the above equations are interpreted as "significantly greater than zero" and "not significantly greater than zero". After all variables of $\mathcal{Z}$ are considered, those that remain in $\mathcal{Z}$ are the direct causal parents of Y, meaning that information flows directly from the elements of $\mathcal{Z}$ to Y. This assertion was proved by the optimal Causation Entropy Principle (oCSE) [40] that provides multiple characterizations of the set of causal parents [37]. The causal parent relationship is written $X_i \to Y$. Note that the collection of all relationships $X_i \to X_j$ forms a directed graph in which the variables are the nodes and edges represent the direction of information flow. This graph can have cycles. In fact it is quite possible that $X_i \to X_j$ and $X_j \to X_i$. In a swarm this situation would be analogous to a "dance" in which two midges were interacting and *mutually* adjusting their movements in accordance with the other insect's movement.

Although CSE is defined theoretically as an integral in the continuous variable case, in practice one must estimate CSE from a finite amount of data samples. Since CSE is expressed as a conditional mutual information, what is needed is essentially a "good" estimator for conditional mutual information. Development of such estimators is an important computational and statistical problem that is of general relevance to a significant body of work in the literature. In this work we adopted a nonparametric estimator of conditional mutual information derived in [41] from the Kraskov-Strögbauer-Grassberger (KSG) [42] estimator for mutual information. Details of this estimator are provided in Appendix A.

3. Application of the oCSE Algorithm to Swarming Insects

Both to demonstrate the types of information that can be gleaned from the direct interaction networks inferred by the oCSE approach and to show that such networks are computable for real empirical data sets that contain noise and other non-idealities, we apply oCSE to empirical measurements of swarming insects. Here, we briefly describe the experimental methodology, including the insect husbandry procedures and data acquisition system, and then show the results of the oCSE computation. These results enable us to compare and contrast spatially nearest neighbors with direct causal neighbors.

A. Experimental Methods

Many different species of insects in the order Diptera exhibit swarming as a part of their mating ritual [43], and such swarms are a well studied, canonical example of collective behavior. Swarms are also an excellent model system for testing the oCSE algorithm: since swarms are internally disordered and

shown little overall pattern or correlation [44], it is difficult to tell by eye which individuals, if any, are interacting.

Here, we apply the oCSE algorithm to data collected from the observation of swarms of the non-biting midge *Chironomus riparius* under controlled laboratory conditions. Details of our insect husbandry procedures and experimental protocols have been reported in detail elsewhere [45], [46], so we described them only briefly here. Our breeding colony of midges is kept in a cubic enclosure measuring 91 cm on a side; temperature and humidity are controlled via laboratory climate-control systems. Midge larvae develop in 9 open tanks, each containing 7 L of oxygenated, dechlorinated water and a cellulose substrate into which the larvae can burrow. Adult midges live in the same enclosure, typically sitting on the floor or walls when they are not swarming. The entire enclosure is illuminated from above by a light source that provides 16 hours of light in each 24-hour period. When the light turns on and off, male midges spontaneously form swarms. We encourage swarm nucleation and position the swarms in the enclosure by means of a "swarm marker" (here, a 32×32 cm piece of black cloth) placed on the floor of the enclosure. The number of midges participating in each swarming event is uncontrolled; we have observed swarms consisting of as few as one or two midges and as many as nearly 100 [47].

To quantify the kinematics of the midges' flight patterns, we reconstruct the time-resolved trajectory of each individual midge via automated optical particle tracking. The midge motion during swarming is recorded by three Point Grey Flea3 digital cameras, which capture 1 megapixel images at a rate of 100 frames per second (fast enough to resolve even the acceleration of the midges [45]). The three cameras are arranged in a horizontal plane outside the midge enclosure with angular separations of roughly 45°. Bright light can disrupt the natural swarming behavior of the midges; thus, we illuminate them in the near infrared, which the midges cannot see but that the cameras can detect. In each 2D image on each camera, midge positions are determined by simple image segmentation followed by the computation of intensity-weighted centroids. These 2D positions were then combined together into 3D world coordinates via stereomatching, using a pinhole model for each camera and calibrating via Tsai's method [48]. To match the individual 3D positions together into trajectories, we used a fully automated predictive particle-tracking method originally developed to study highly turbulent fluid flows [49]. Occasionally, tracks will be broken into partial segments, due to mistakes in stereoimaging or ambiguities in tracking; to join these segments together into long trajectories, we used Xu's method of re-tracking in a six-dimensional position-velocity space [50]. After tracks were constructed, accurate velocities and accelerations were computed by convolving the trajectories with a smoothing and differentiating kernel [47]. The final data set for each swarming event therefore consists of time series of the 3D position and its time derivatives for each midge.

B. Results

The most basic result of applying the oCSE algorithm is the determination of the *direct* causal links between individuals in the swarm. Although the data contains many suitable time series describing the motion of more than 30 swarming midges, Figs. 3 and 4 describe a small swarm of 5 midges for the purpose of illustrating the application of CSE to finding direct causal links. In Fig. 3, we show four consecutive snapshots of these links from the perspective of a single insect (labelled as "1") over a period of 2 seconds. In the first snapshot, panel (a), midge 1 is identified as being influenced by midges 2 and 4; that is, it is receiving information from them. Notice that in the second snapshot, (b), the link from 2 to 1 has been lost, but 1 is still receiving information from 4. Perhaps because it moved closer, 1 is also receiving information from 3 in the second snapshot. By panel (c), 1 seems to have noticed 5, but by the final snapshot this link has been lost.

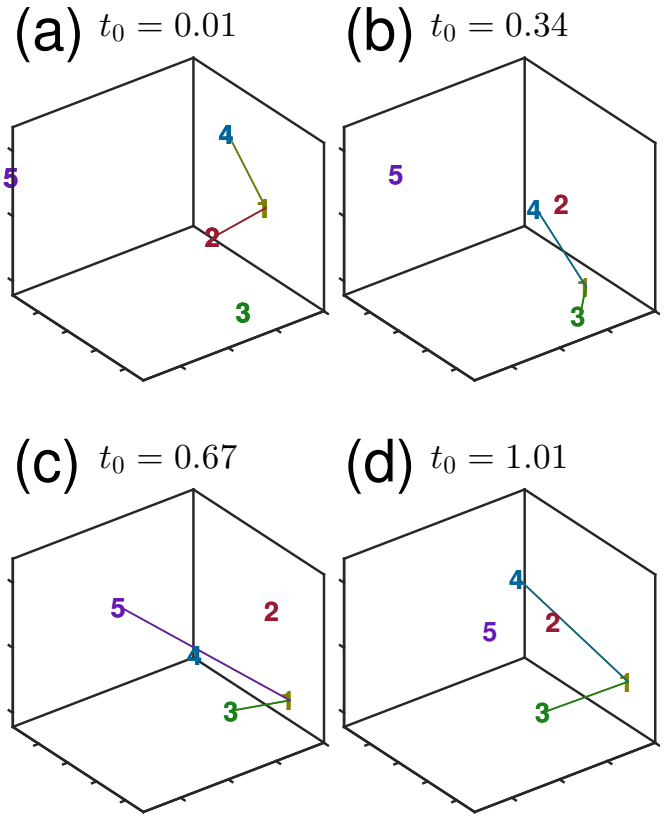

Fig. 3. Information flow from the perspective of midge 1. Only information flows into midge 1 (i.e., direct causal links to 1) are depicted, and are represented by the solid lines. Each panel corresponds to an oCSE computation using 1 second (100 frames) of data. The positions of the midges are given by their initial positions, t_0 during the interval. Direction in time is created by lagging variables by $\tau = 0.05$ seconds. The initial time $t_0 = 0.01$ of panel (a) is chosen to be the point when insect 5 becomes observable.

Such transient interactions are reminiscent of those we described earlier using a different (time-series-based) measure [19]. In that case, we had hypothesized that the primary purpose of such interactions was for the registration of the gender of other midges in the swarm, since the biological purpose of swarming in this species is mating. A similar process may be at work here, and midge 1 may have, for example, successfully identified midge 2 after the first snapshot so that further information transfer was unnecessary.

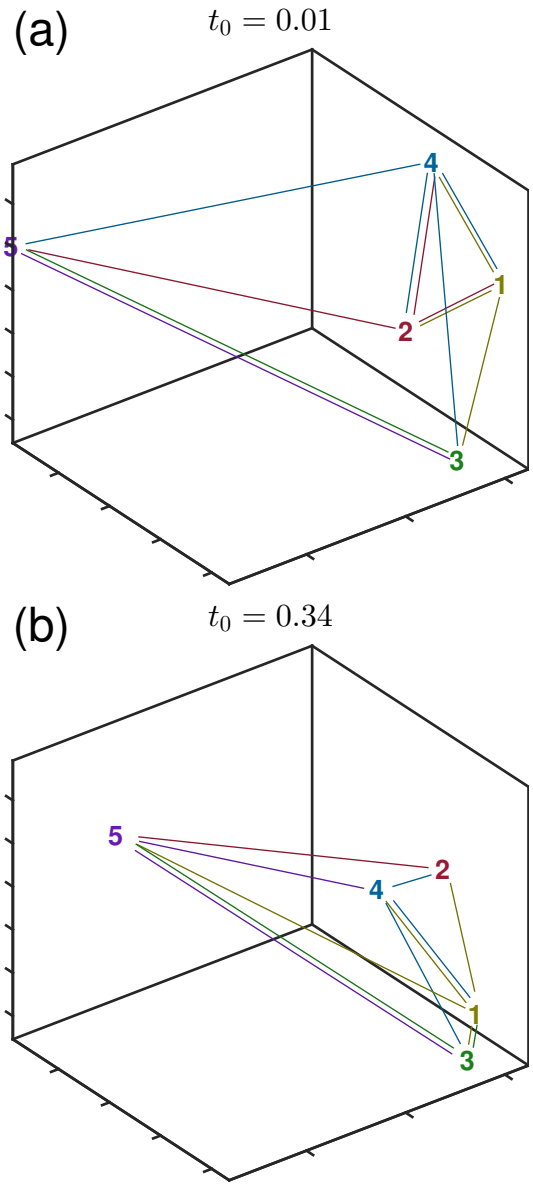

Fig. 4. Directed graphs of all inferred information flows corresponding to the first two panels of Fig. 3. Each edge represents a flow of information and takes on the color of the information source. The parameters used in the oCSE algorithm are the same as in Fig. 3: $\tau = 0.05$, each computation uses 100 frames of data, and the positions are determined by the first frame.

In addition to studying only the information flows into a single insect, we can look at the entire directed graph returned by the oCSE algorithm. In Fig. 4, we show this full directed graph for the two initial conditions corresponding to frames (a) and (b) of Fig. 3. The direction of an edge is given by its color, where the source of the information flow determines the color.

One easy set of statistics to read off of these graphs are the in and out degrees. The in degree of a node is the number of edges pointing to that node and the out degree is the number of edges with one end at the node but pointing to a different node. In these plots, then, the out degree of a node is the number of edges that are plotted in the same color as the node and the in degree is the number of edges with an end at the node but which are plotted in a different color. So, for instance, in Fig. (4a), the in degree of midge 1 is 2 (verifying the computation used to create Fig. (3a)), and the out degree is 3.

The average in degree (which is always equal to the average out degree) is $12/5 = 2.4$ in both (a) and (b) of Fig. 4, suggesting that on average the number of other midges that any given midge is paying attention to may be relatively constant in time. This theory is lent more credence by noting that the in degree of every individual midge is the same in (a) and (b).

The out-degrees are much more variable, however. Biologically speaking, out degrees may give information on which midges are the most important, in the sense that if a midge has a high out degree then others seem to be reacting to the motions of this midge. In Fig. (4a), although the most spatially central node, midge 2, has an out-degree of 3, so does midge 1, which is not as spatially central. Furthermore, midge 4 has the largest out degree with every other midge paying attention to 4. So, although midge 2 is the most spatially central node, we say that midge 4 has the highest "degree centrality". A similar analysis can be carried out on panel (b) showing that at $t_0 = 0.34$, midge 4 is now the most spatially central, but node 1 has the highest degree centrality. Some statistics that give more detailed information about centrality are eigenvector centrality and betweenness centrality.

Rather than attempting to comprehensively apply all available graph analysis methods, we give two simple observations and refer the reader to [51] for more on the analysis of graphs. In Fig. (4a) the subset of midges $\{3, 5\}$ forms a "sink" for information. Although 3 and 5 are gathering information from many midges, the are apparently unable to send that information to any other midges than 3 and 5. If one were to code edges in an "adjacency matrix" of 1's representing edges and 0's represent the lack of an edge, this feature corresponds to the adjacency matrix being reducible. The set $\{3, 5\}$ generates the only non-trivial subgraph closed under inclusion of all out-going edges. A less restrictive analysis that is similar in flavor is community detection [51], but this type of analysis is usually reserved for larger graphs.

Again in Fig. (4a), the edges linking $\{1, 2, 4\}$ generate a triangle in which information can flow in both directions around the triangle. This is a special relationship between three nodes called a 3-clique. It should be compared with $\{2, 4, 5\}$ in Fig. (4) in which information flows only in one direction. In swarms with many other midges it is conceivable that most randomly picked triplets $\{a, b, c\}$ would have no triangle between them. The density of different types of triangles in a set is quantified by the clustering coefficient. Social networks tend to have much higher clustering as measured by clustering coefficients than technological networks and networks whos edges are determined randomly [52].

Because both Figs. 3 and 4 show that the configuration of causal links can and does change in time, it is reasonable to ask about the temporal variability and stability of the oCSE results: if the links switch seemingly at random from time step to time step, then the results would be unintelligibly unreliable. To check the stability and reliability of the oCSE

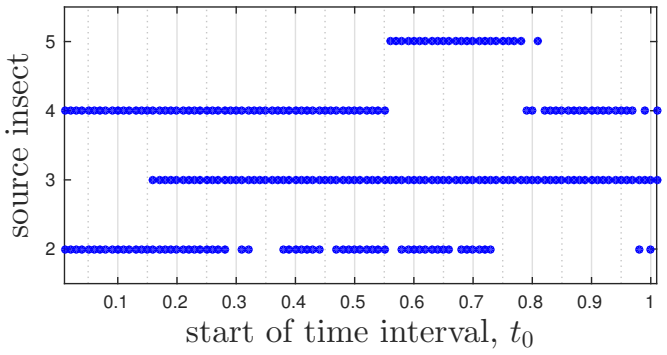

Fig. 5. Evolution of a causal neighborhood of a single midge over one second following the trajectories of 5 midges. This figure gives a more (temporally) detailed look at the information flows depicted in Fig. 3 at the expense of spatial information. Again, midge 1 is the target individual; that is, we are inferring causal links toward midge 1. Edges from j to 1 are replaced by a point at (t_0, j) for the appropriate t_0. The inputs and parameters are the same as those used to create Fig. 3. Long stretches of symbols or blank space demonstrate that the oCSE algorithm is robust to changes in the spatial configuration and kinematics of the swarm.

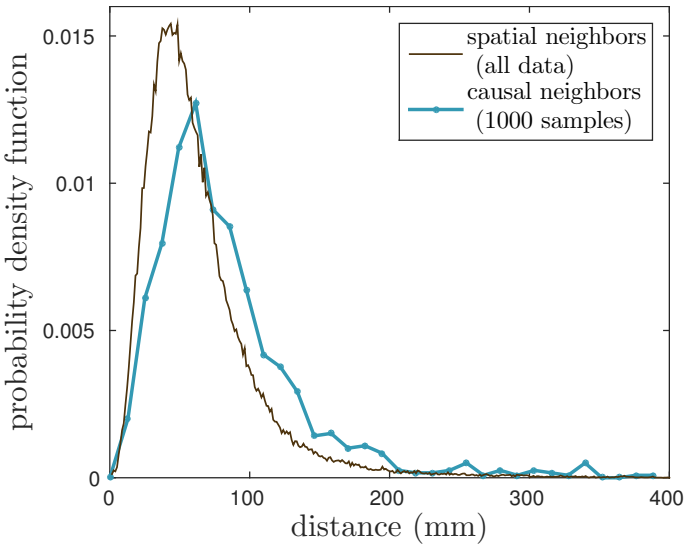

Fig. 6. Probability density functions for distance from an arbitrary midge to its nearest spatial neighbor and distance to its nearest causal neighbor. Both pdfs are estimated by binning sets of distances. The distances used to form the nearest neighbor distribution were calculated by sampling a random midge from each time slice such that the time slice is included in a time interval that contains a constant number of insects that is greater than or equal to 5 and calculating the minimum of the distances to all other midges. There were greater than 145k such times slices. The distances to nearest causal neighbors were formed by randomly sampling these 145k time slices 1000 times in such a way that the $T = 1$ second time intervals involved in the CSE computations would not overlap. The number of midges per sample ranged from 5 to 30. Again a random insect was chosen, but oCSE was used to compute the set of causal neighbors and the minimum distance chosen from the distances of these neighbors to the target. The parameters for the oCSE algorithm, $T = 1$ second and $\tau = 0.05$ seconds were the same as those used in previous figures. There were three time slices that produced empty causal structures. Such time slices were discarded.

results, we computed the causal links for sets of overlapping time intervals, as shown in Fig. 5 for a particular example. Although there is occasionally some drop-out of links from one instant to the next, the overall results of the algorithm are clearly stable and more-or-less continuous in time; and we conjecture that the discontinuously dropped causal links could in fact be restored, if necessary, by improving the tracking accuracy or by some post-processing step or some combination of both.

Finally, as noted above, it is intriguing to note that midges connected via causal links are not always the spatially closest to each other. Although a full characterization and complete understanding of the distinction between spatial proximity and causal information flow is beyond the scope of the present paper, we can at least describe at a statistical and macroscopic level the difference between those by measuring the probability density functions (pdfs) of the distance between nearest spatial neighbors and nearest causal neighbors. These pdfs are plotted in Fig. 6 and show that as compared to simply calculating minimum distances for given time slices from the raw data, the restriction to causal neighbors by the oCSE algorithm shifts the typical distance between neighbors to larger values. The average distance from an insect to its nearest spatial neighbors and to its nearest causal parent are estimated to be 63.1 and 82.8 mm respectively, with corresponding standard errors (i.e. standard deviations of the estimate of the mean) of 0.1 and 1.7 mm.

4. Conclusion

Information theoretic methods show great promise for the analysis of biological data. In particular, the oCSE algorithm allows the inference of directed graphs of information flow from time series data. The introduction of the oCSE algorithm allows a wider variety of questions about collective animal motion to be addressed. In particular, this allows us to pose the question of which individual animals are *directly* interacting with which other individual animals. Such analysis opens up possibilities and questions well beyond standard group analysis, which may typically be based on mean-field behaviors. Specifically, here we have suggested that it is possible to consider which animal may be acting as a center of attention, and how this scenario may change in a time-varying network of information flow of influence. Further we have noted that contrary to standard models of swarming and group behaviors, these information flow networks allow that the influential animals for each individual animal may not necessarily be the closest spatial neighbors. With these new analysis tools, we hope that a wide variety of new questions can be posed and studied directly from experimental data, moving beyond current phenomenological models that are validated in terms of interpretive rather than observational understanding. Without being able to anticipate all the future questions that can be discussed within this paradigm of interaction being defined by direct information flow as CSE, we hope that it is clear that this new tool and this new perspective allows for a bright future for experimental sciences in a wide variety of studies in collective animal behaviors.

Appendix A
Estimation of CSE from Data

In practice, formula (2.5) defining CSE must be estimated from time series data. By time lagging the time series for the

effect variable, the estimation of $C_{X\to Y|Z}$ can be reduced to the estimation of the conditional mutual information, or to the estimation of mutual information when the conditioning set is empty.

A straightforward way to estimate entropy and related quantities is by binning (histogram) methods, which effectively estimate a probability density $p(x)$ by counting the frequency of sampled points falling in a constant-size region around that point. Although conceptually simple to understand and easy to implement, such binning methods have been shown to suffer from slow convergence, especially for multivariate data sets, as they scale badly with the embedding dimension. Faster convergence can be achieved by nonparametric estimators based on k-nearest neighbor (knn) statistics. The basic idea is to estimate the density at a given point using distance to the k nearest neighbors rather than neighbors falling within a constant-size neighborhood. For mutual information, we adopt the Kraskov-Strögbauer-Grassberger (KSG) estimator, which was shown to be data efficient (with $k = 1$ the estimator resolve structures down to the smallest possible scales), adaptive (the resolution is higher where data are more numerous), and to have minimal bias (the bias is mainly due to nonuniformity of the density at the smallest resolved scale, giving typical systematic errors that scale as functions of k/N for N points) [42]. The KSG estimator can be extended for the estimation of conditional mutual information. One such extension was recently proposed by Vejmelka and Palus (VP) [41]. Consider n independent samples $\{w_1, w_2, , \ldots, w_n\}$ of the joint random variable $W = (X, Y, Z)$ where $w_i = (x_i, y_i, z_i)$. The VP estimator of $I(X;Y|Z)$ is given by

$$\hat{I}^{(VP)}(X;Y|Z) = \psi(k) - \langle \psi(n_{xz}+1) + \psi(n_{yz}+1) - \psi(n_z+1) \rangle. \tag{A.1}$$

Here $\langle\cdot\rangle$ denotes the average over the samples and $\psi(t) = \Gamma'(t)/\Gamma(t)$ is the digamma function. For fixed value of k, let $\epsilon(i)$ denotes the distance from $w_i \equiv (x_i, y_i, z_i)$ to its kth nearest neighbor in the joint space (x, y, z), from which

- $n_{xz}(i)$ is the number of points (x_j, z_j) $(j \neq i)$ with $\|(x_j, z_j) - (x_i, z_i)\| < \epsilon(i)$
- $n_{yz}(i)$ is the number of points (y_j, z_j) $(j \neq i)$ with $\|(y_j, z_j) - (y_i, z_i)\| < \epsilon(i)$
- $n_z(i)$ is the number of points z_j $(j \neq i)$ with $\|z_j - z_i\| < \epsilon(i)$

Several recent papers have focused on reducing the finite-sample bias either of the KSG-VP type of estimators [53]–[55] or other types of estimators [56].

Appendix B
Significance Test of CSE

A key step in the algorithmic inference of direct causal links from data is determining whether or not the estimated CSE value $\widehat{C_{X\to Y|Z}}$ should be regarded as being strictly positive. To address this question, we consider a shuffle test for the null hypothesis of $\widehat{C_{X\to Y|Z}} = 0$ [37]. Given time series samples $\{(x_t, y_t, z_t)\}$ of a stochastic process (X_t, Y_t, Z_t), the idea is that for the estimated $\widehat{C_{X\to Y|Z}} > 0$ to be regarded as significant, it should typically "beat" (i.e., be greater than) $\widehat{C_{X'\to Y|Z}}$ where X' corresponds to a surrogate data obtained by replacing $\{x_t\}$ with $\{x'_t\}$, the latter being a random permutation of the set $\{x_t\}$. By repeating the estimation with a large number of permutations, we consider $C_{X\to Y|Z}$ to be significant if $\widehat{C_{X\to Y|Z}}$ is greater than a fraction $(1-\alpha)$ of the values of $\widehat{C_{X'\to Y|Z}}$ where α is a prescribed level of significance.

Appendix C
Parameters Choice for the Network Construction using oCSE

The data set considered herein has consisted of 126 time series representing the positions and accelerations of varying sized swarms of insects for different lengths of time. We narrowed down the data by searching for long time intervals in which contained exactly the same number of insects for a relatively long period of time. In other words we restricted our studies to data sets where the same "actors" were at play throughout the window of study. We used the acceleration data for the computations of CSE because the acceleration showed less autocorrelation than the position data.

The KSG and VP estimators of mutual information require a choice of parameter k, the number of neighbors defining a neighborhood in the joint space. We chose $k = 4$ for all computations. This choice of k is arbitrary except for wanting to avoid very small k, which give estimates with high variance. A number of papers offer heuristics for choosing k [57], [58] as a function of sample size.

We chose hypothesis tests with significance level $\alpha = 0.01$ in both the forward and backward steps of the oCSE algorithm. In theory, α should control the sparsity of the desired graph but we were unable to confirm this numerically. We typically ran 1000 trials per hypothesis test.

Acknowledgment

This work was funded by the Army Research Office (ARO) under Grant No. W911NF-12-1-0276 (WML, JS, EMB), Grant No. W911NF-16-1-0081 (WML, JS, EMB), Grant No. W911NF-16-1-0185 (NTO), and the Simons Foundation under Grant No. 318812 (JS). We thank Dr. Samuel Stanton from the ARO Complex Dynamics and Systems Program for his ongoing and continuous support.

References

[1] L. Spector, J. Klein, C. Perry, and M. Feinstein, "Emergence of collective behavior in evolving populations of flying agents," *Genetic Programming and Evolvable Machines*, vol. 6, pp. 111–125, 2005.

[2] A. Sokolov, I. S. Aranson, J. O. Kessler, and R. E. Goldstein, "Concentration dependence of the collective dynamics of swimming bacteria," *Phys. Rev. Lett.*, vol. 98, p. 158102, 2007.

[3] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 4, pp. 1232–1237, 2008.

[4] E. Barreto, B. Hunt, E. Ott, and P. So, "Synchronization in networks of networks: The onset of coherent collective behavior in systems of interacting populations of heterogeneous oscillators," *Phys. Rev. E*, vol. 77, p. 036107, 2008.

[5] I. Giardina, "Collective behavior in animal groups: theoretical models and empirical studies," *HFSP Journal*, vol. 2, no. 4, pp. 205–219, 2008.

[6] T. S. Deisboeck and I. D. Couzin, "Collective behavior in cancer cell populations," *BioEssays*, vol. 31, pp. 190–197, 2009.

[7] M. Moussaid, S. Garnier, G. Theraulaz, and D. Helbing, "Collective information processing and pattern formation in swarms, flocks, and crowds," *Topics in Cognitive Science*, vol. 1, pp. 469–497, 2009.

[8] C. A. Yatesa, R. Erban, C. Escudero, I. D. Couzind, J. Buhl, I. G. Kevrekidis, P. K. Maini, and D. J. T. Sumpter, "Inherent noise can facilitate coherence in collective swarm motion," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 14, pp. 5464–5469, 2009.

[9] T. Gregor, K. Fujimoto, N. Masaki, and S. Sawai, "The onset of collective behavior in social amoebae," *Science*, vol. 328, no. 5981, pp. 1021–1025, 2010.

[10] R. Lukemana, Y.-X. Li, and L. Edelstein-Keshet, "Inferring individual rules from collective behavior," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 28, pp. 12 576–12 580, 2010.

[11] D. J. T. Sumpter, *Collective Animal Behavior*. Princeton University Press, 2010.

[12] H. P. Zhang, A. Be'er, E.-L. Florin, and H. L. Swinney, "Collective motion and density fluctuations in bacterial colonies," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 31, pp. 13 626–13 630, 2010.

[13] M. Hentschel, D. Dregely, R. Vogelgesang, H. Giessen, and N. Liu, "Plasmonic oligomers: The role of individual particles in collective behavior," *ACS Nano*, vol. 5, no. 3, pp. 2042–2050, 2011.

[14] M. Moussaïda, D. Helbing, and G. Theraulaz, "How simple rules determine pedestrian behavior and crowd disasters," *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 17, pp. 6884–6888, 2011.

[15] B. Ravoori, A. B. Cohen, J. Sun, A. E. Motter, T. E. Murphy, and R. Roy, "Robustness of optimal synchronization in real networks," *Phys. Rev. Lett.*, vol. 107, p. 034102, 2011.

[16] J. Sun, E. M. Bollt, M. A. Porter, and M. S. Dawkins, "A mathematical model for the dynamics and synchronization of cows." *Physica D*, vol. 240, pp. 1497–1509, 2011.

[17] P. S. Skardal, D. Taylor, and J. Sun, "Optimal synchronization of complex networks," *Phys. Rev. Lett.*, vol. 113, p. 144101, 2014.

[18] R. Ni, J. G. Puckett, E. R. Dufresne, and N. T. Ouellette, "Intrinsic fluctuations and driven response of insect swarms," *Phys. Rev. Lett.*, vol. 115, p. 118104, 2015.

[19] J. G. Puckett, R. Ni, and N. T. Ouellette, "Time-frequency analysis reveals pairwise interactions in insect swarms," *Phys. Rev. Lett.*, vol. 114, p. 258103, 2015.

[20] J. K. Parrish and L. Edelstein-Keshet, "Complexity, pattern, and evolutionary trade-offs in animal aggregation," *Science*, vol. 284, pp. 99–101, 1999.

[21] M. Tennenbaum, Z. Liu, D. Hu, and A. Fernandez-Nieves, "Mechanics of fire ant aggregations," *Nat. Mater.*, vol. 15, pp. 54–59, 2016.

[22] R. Ni and N. T. Ouellette, "On the tensile strength of insect swarms," *submitted*, 2016.

[23] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, "Inferring individual rules from collective behavior," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 12 576–12 580, 2010.

[24] J. E. Herbert-Read, A. Perna, R. P. Mann, T. M. Schaerf, D. J. T. Sumpter, and A. J. W. Ward, "Inferring the rules of interaction of shoaling fish," *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 18 726–18 731, 2011.

[25] Y. Katz, K. Tunstrom, C. C. Ioannou, C. Huepe, and I. D. Couzin, "Inferring the structure and dynamics of interactions in schooling fish," *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 18 720–18 725, 2011.

[26] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, "Novel type of phase transition in a system of self-driven particles," *Phys. Rev. Lett.*, vol. 75, pp. 1226–1229, 1995.

[27] C. W. Reynolds, "Flocks, herds, and schools: A distributed behavioral model," *SIGGRAPH Comput. Graph.*, vol. 21, pp. 25–34, 1987.

[28] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, "Collective memory and spatial sorting in animal groups," *J. Theor. Biol.*, vol. 218, pp. 1–11, 2002.

[29] J. Toner, Y. Tu, and S. Ramaswamy, "Hydrodynamics and phases of flocks," *Ann. Phys.*, vol. 318, pp. 170–244, 2005.

[30] N. T. Ouellette, "Empirical questions for collective-behaviour modelling," *Pramana-J. Phys.*, vol. 84, pp. 353–363, 2015.

[31] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[32] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, p. 845, 2000.

[33] N. W. F. Bode, A. J. Wood, and D. W. Franks, "The impact of social networks on animal collective motion," *Anim. Behav.*, vol. 82, pp. 29–38, 2011.

[34] ——, "Social networks and models for collective motion in animals," *Behav. Ecol. Sociobiol.*, vol. 65, pp. 117–130, 2011.

[35] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.85.461

[36] J. Sun and E. M. Bollt, "Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings." *Physica D*, vol. 267, pp. 49–57, 2014.

[37] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM J. Appl. Dyn. Syst.*, vol. 14, no. 1, pp. 73–106, 2015.

[38] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2006.

[40] Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM J Appl Dyn Syst*, vol. 14, no. 1, pp. 73–106, 2015.

[41] M. Vejmelka, M. Paluš, and Milan, "Inferring the directionality of coupling with conditional mutual information," *Phys. Rev. E*, vol. 88, p. 026214, Feb 2008. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.77.026214

[42] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.69.066138

[43] J. A. Downes, "The swarming and mating flight of *Diptera*," *Annu. Rev. Entomol.*, vol. 14, pp. 271–298, 1969.

[44] R. Ni and N. T. Ouellette, "Velocity correlations in laboratory insect swarms," *Eur. Phys. J. Special Topics*, vol. 224, pp. 3271–3277, 2015.

[45] D. H. Kelley and N. T. Ouellette, "Emergent dynamics of laboratory insect swarms," *Sci. Rep.*, vol. 3, p. 1073, 2013.

[46] J. G. Puckett, D. H. Kelley, and N. T. Ouellette, "Searching for effective forces in laboratory insect swarms," *Sci. Rep.*, vol. 4, p. 4766, 2014.

[47] J. G. Puckett and N. T. Ouellette, "Determining asymptotically large population sizes in insect swarms," *J. R. Soc. Interface*, vol. 11, p. 20140710, 2014.

[48] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robotic. Autom.*, vol. RA-3, pp. 323–344, 1987.

[49] N. T. Ouellette, H. Xu, and E. Bodenschatz, "A quantitative study of three-dimensional Lagrangian particle tracking algorithms," *Exp. Fluids*, vol. 40, pp. 301–313, 2006.

[50] H. Xu, "Tracking Lagrangian trajectories in position-velocity space," *Meas. Sci. Technol.*, vol. 19, p. 075105, 2008.

[51] M. E. J. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[52] ——, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003. [Online]. Available: http://dx.doi.org/10.1137/S003614450342480

[53] S. Gao, G. V. Steeg, and A. Galstyan, "Estimating mutual information by local gaussian approximation," *arXiv preprint arXiv:1508.00536*, 2015.

[54] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1 – 46, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0370157307000403

[55] J. Zhu, J.-J. Bellanger, H. Shu, C. Yang, and R. Le Bouquin Jeannès, "Bias reduction in the estimation of mutual information," *Phys. Rev. E*, vol. 90, p. 052714, Nov 2014. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.90.052714

[56] M. T. Giraudo, L. Sacerdote, and R. Sirovich, "Non–parametric estimation of mutual information through the entropy of the linkage," *Entropy*, vol. 15, no. 12, p. 5154, 2013. [Online]. Available: http://www.mdpi.com/1099-4300/15/12/5154

[57] Y. Mack and M. Rosenblatt, "Multivariate k-nearest neighbor density estimates," *Journal of Multivariate Analysis*, vol. 9, no. 1, pp. 1 – 15, 1979.

[58] K. Sricharan, R. Raich, and A. O. H. III, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Information Theory*, vol. 58, no. 7, pp. 4135–4159, 2012. [Online]. Available: http://dx.doi.org/10.1109/TIT.2012.2195549