Topological analysis of complexity in multiagent systems

Nicole Abaid,[1] Erik Bollt,[2] and Maurizio Porfiri[1,*]

[1] *Department of Mechanical and Aerospace Engineering, Polytechnic Institute of New York University, Brooklyn, New York 11201, USA*

[2] *Department of Mathematics, Clarkson University, Potsdam, New York 13699, USA*

(Received 26 July 2011; revised manuscript received 19 December 2011; published 10 April 2012)

Social organisms at every level of evolutionary complexity live in groups, such as fish schools, locust swarms, and bird flocks. The complex exchange of multifaceted information across group members may result in a spectrum of salient spatiotemporal patterns characterizing collective behaviors. While instances of collective behavior in animal groups are readily identifiable by trained and untrained observers, a working definition to distinguish these patterns from raw data is not yet established. In this work, we define collective behavior as a manifestation of low-dimensional manifolds in the group motion and we quantify the complexity of such behaviors through the dimensionality of these structures. We demonstrate this definition using the ISOMAP algorithm, a data-driven machine learning algorithm for dimensionality reduction originally formulated in the context of image processing. We apply the ISOMAP algorithm to data from an interacting self-propelled particle model with additive noise, whose parameters are selected to exhibit different behavioral modalities, and from a video of a live fish school. Based on simulations of such model, we find that increasing noise in the system of particles corresponds to increasing the dimensionality of the structures underlying their motion. These low-dimensional structures are absent in simulations where particles do not interact. Applying the ISOMAP algorithm to fish school data, we identify similar low-dimensional structures, which may act as quantitative evidence for order inherent in collective behavior of animal groups. These results offer an unambiguous method for measuring order in data from large-scale biological systems and confirm the emergence of collective behavior in an applicable mathematical model, thus demonstrating that such models are capable of capturing phenomena observed in animal groups.

DOI: 10.1103/PhysRevE.85.041907 PACS number(s): 87.23.Cc, 89.75.Fb, 02.70.−c, 05.65.+b

I. INTRODUCTION

Schooling of fish [1], swarming of bacteria [2], and flocking of birds [3] are all instances of collective behavior of biological groups [4]. Such groups typically comprise a large number of individuals, whose motion involves complex body deformations and harmonious speed changes. In fish schools, collective behavior is identified by characteristic spatiotemporal patterns, including relatively small distances among adjacent individuals, hydrodynamically motivated staggering of leaders and followers, and sharp delineation of the group as a whole in space [1,5]. Although every motivation of fish schooling is not yet identified, members of a school benefit from many aspects of social life, such as increased predator evasion, foraging capabilities, and swimming efficiency [6,7].

Fish schools may be extremely varied, bringing together individuals of different species [8,9], temperaments [10], knowledge [11–13], and may include widely different cardinalities [14]. Beyond the composition of the animal group, environmental stimuli, such as perceived predation [15] and light intensity [16,17], can promote collective versus individual behavior. One remarkable facet of a fish school is that its complex structure emerges and is maintained by the exchange of multifaceted information among peers. In other words, fish schools persist without the presence of a permanent leader [11]; instead, the structure is based on local decisions made by individuals [7].

Understanding the behaviors driving such patterns is currently the subject of intensive biology and mathematical research [18–26]. Besides its clear applications in the biological disciplines, a fuller comprehension of such complex systems has the potential to inform the design of cooperative control algorithms for multivehicle teams [27], data fusion methods for wireless sensor networks [28], and intelligent power distribution systems [29]. Even if the data describing such ubiquitous phenomena across different time scales and spatial lengths may be huge, recognizing the emergence of collective behavior is a surprisingly simple task routinely executed by trained or untrained observers [30]. Motivated by this ability of human cognition, we consider such simplicity to be a manifestation of inherently low-dimensional structures underlying large scale systems. A similar simplification has been used to characterize the motion of nematodes in [31] by classifying body attitudes as combinations of a small number of characteristic postures. In this work, we take a different approach as we completely remove the human observer and directly apply a dimensionality reduction algorithm to large scale data sets.

Specifically, we propose a formal definition of collective behavior as the existence of a low-dimensional embedding stable invariant manifold in the full space of the trajectories of the agents comprising the group. Existence of a stable invariant manifold usually is due to dissipation in the system. As a complementary characterization of this notion of collective behavior, we define the relative dimension of the minimal embedding manifold as the degree of complexity of the system. This approach seeks to eliminate ansatz on order parameters for measuring collective behavior by using exclusively raw data pertaining to the manifestation of the phenomenon.

[*] mporfiri@poly.edu

In turn, such objective definition may allow for the direct validation of order parameters and eventually inform their formulation.

As a test bed of collectively acting agents, we focus on so-called Vicsek biological groups modeled as interacting self-propelled particles in a discrete-time setting [32]. This modeling framework offers a flexible platform for exploring different phenomena taking place in animal groups, such as long-range attraction, short-range repulsion, and perceptual limitation [3,19,33–36]. With reference to fish schooling, this model implements a consensus protocol for the individuals' headings, from which complex macroscopic behaviors emerge. By varying the model parameters, we induce biologically- relevant behaviors such as highly aligned regular motion featuring group mates swimming along the same direction with constant speed and regular milling about a fixed location in a mobbing-type behavior. Both these maneuvers are executed by live fish schools [1,37]. As a numerical control experiment, we also consider the trivial model which prohibits interaction among agents and thus relegates them to be random walkers.

We test the working definition of collective behavior using the recently developed isometric mapping (ISOMAP) algorithm, which offers a simplified perspective of large scale data sets by embedding such data on lower-dimensional manifolds [38–40]. Manifold learning is an established and rapidly evolving area in the machine learning community for a wide variety of practical problems which require detecting low-dimensional structures in very high-dimensional data sets. For example, it is ideal for the classification and feature extraction problems of handwritten character recognition [41], object recognition [42], and facial recognition [42]. There are a multitude of popular algorithms and methods of manifold learning from data [43]. Among these approaches, the ISOMAP algorithm offers a viable solution for data-driven analysis of dynamical systems by characterizing low-dimensional invariant manifolds therein, as demonstrated in [44]. This algorithm, originally formulated in the context of data mining and image processing, approximates an invariant manifold by an undirected graph whose geodesics coincide with those of the true nonlinear manifold. This perspective bears some resemblance in spirit to the literature on the theory of time delay embedding [45]. However, the ISOMAP algorithm allows for a global model of an invariant manifold as an undirected graph which preserves distances along the manifold. Beyond its technical sophistication, the ISOMAP algorithm is easy to implement and several software suites are readily available for use [46].

In this work, we employ the ISOMAP algorithm to illustrate our working definition of collective behavior using both simulation data from the self-propelled particle model and raw image data of a live fish school. As a stand-in for the human observer, we identify complexity in simulation data using traditional, behaviorally defined measures of alignment and cohesion among individuals. Corroborating traditional measures, we find that increasing complexity in group behavior corresponds to an increasing dimensionality of the minimal embedding manifolds. In addition, such low-dimensional manifolds are entirely absent from the trivial model wherein agents are random walkers, thus confirming the validity of this definition of collective behavior.

II. SELF-PROPELLED PARTICLE MODEL

We consider a system of N interacting agents traveling in a two-dimensional domain at a speed s. We take a square domain of side length $L \in \mathbb{R}^+$ and we select reflective boundary conditions. The two-dimensional position of the agents at time $k \in \mathbb{Z}^+$ is given by the vector $x(k) \in \mathbb{C}^N$, where the real and imaginary parts of the ith element $x_i(k)$ belong to $[-L/2, L/2]$ for $i = 1, \dots, N$. At time k, agent i has heading denoted $\theta_i(k) \in [-\pi, \pi]$, where $\theta_i(k) = 0$ corresponds to heading along the positive real axis. The agents have uniformly distributed random initial conditions for both position and heading.

While agent i maintains a constant speed at any time, it progressively updates its heading according to the interactions with neighbors. Specifically, at time k, agent i interacts with all agents j, for $j = 1, \dots, N$, such that $\|x_j(k) - x_i(k)\| \leqslant r$, where $\|\cdot\|$ is the Euclidean norm and $r \in \mathbb{R}^+$ is constant. We use the notation $\mathcal{N}_i(k)$ to identify such neighbors. The presence of an exogenous stimulus is modeled by including a source at $x_0 \in \mathbb{C}^N$ which, at time k, acts on agent i's heading update as a virtual neighbor when $\|x_0 - x_i(k)\| \leqslant r_a$ for $r_a \in \mathbb{R}^+$. We refer to $\mathcal{N}_0(k)$ as the neighbor set of this source at time k.

At successive time steps, agent i updates its heading according to

$$\theta_i(k+1) = \widehat{\theta}_i(k+1) + P, \tag{1}$$

with $\widehat{\theta}_i(k+1)$ being the consensus-driven heading given by

$$\arg\left(\sum_{j \in \{i\} \cup \mathcal{N}_i(k)} \exp[\iota\theta_j(k)] + \mathrm{ind}_{\mathcal{N}_0(k)}(i) \exp[\iota\phi_i(k)] \right) + \Delta\theta. \tag{2}$$

Here, ι is the imaginary unit, $\mathrm{ind}_{\mathcal{N}_0(k)}(\cdot)$ is the indicator function for $\mathcal{N}_0(k)$, $\phi_i(k) = \arg[x_0 - x_i(k)]$, and

$$P = \begin{cases} p\pi, & \text{for } \widehat{\theta}_i(k+1) \in (-\pi/2, 0] \cup (\pi/2, \pi], \\ -p\pi, & \text{for } \widehat{\theta}_i(k+1) \in (0, \pi/2] \cup [-\pi, -\pi/2]. \end{cases} \tag{3}$$

The quantity $\Delta\theta$ is a uniformly distributed random variable which takes values in $[-\eta\pi, \eta\pi]$ and $\eta \in [0,1]$ is the so-called noise parameter. When $\eta = 0$, the agents reach and maintain a common heading. For η large, the heading of the agents is random at each time step. The parameter p biases the agents' headings to preferentially induce motion parallel to the real axis and, for example, describes fish locomotion parallel to a flow. Agent i updates its position according to

$$x_i(k+1) = x_i(k) + s \, \exp[\iota\theta_i(k+1)]\,. \tag{4}$$

When an agent's updated position exceeds the boundary of the spatial domain, its previous position is maintained and the component of its two-dimensional velocity corresponding to the coordinate violating the spatial boundary is multiplied by -1. As a result, an agent's speed is less than s in instances when the agent encounters the boundary.

We consider two qualitatively different sets of simulation parameters to investigate this model. The first parameter set, which we refer to as "aligned pacing" (AP), uses $p = 0.01$ and no source, that is $\mathcal{N}_0(k) = \emptyset$ for all times k. Using this parameter selection, agents are capable of forming an aligned group which travels along the preferred domain dimension

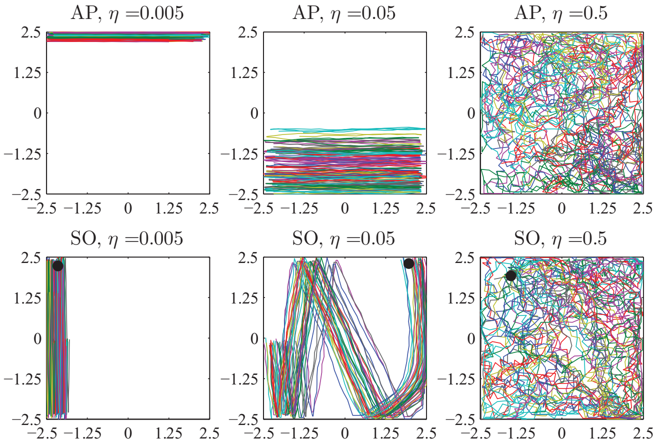

FIG. 1. (Color online) Sample model trajectories with AP and SO parameters and $\eta = 0.005$, 0.05, and 0.5. We define $N = 40$ agents traveling in the complex plane with $s = 0.25$, $L = 5$, $r = 1$, and $r_a = 3$ and we display the first 75 time steps out of a 10 000 time step simulation of the model. For AP simulation data, $p = 0.01$. For SO simulation data, the source position is $-2.15 + 2.23\iota$ when $\eta = 0.005$, $1.96 + 2.30\iota$ when $\eta = 0.05$, and $-1.56 + 1.92\iota$ when $\eta = 0.5$. Horizontal and vertical axes refer to the real and imaginary parts of $x(k)$, respectively, for different values of k.

depending solely on the noise η. The second parameter set, which we call "source orbit" (SO), uses $p = 0$ and a randomly placed source x_0. Setting $p = 0$ allows the interaction with the source, neighbors, and additive noise to determine the agents' spatial positions with no preferred group orientation. Simulation data from this model approximate the behavior of fish schools milling around a common center to avoid predation or to aggregate near an attracting stimulus [37].

Truncated trajectories of $N = 40$ agents interacting according to this model with both parameter sets and representative values of η are presented in Fig. 1. We consider agents with constant speed $s = 0.25$ interacting for 10 000 time steps after omitting 1000 initial time steps to eliminate any transient. We use the simulation parameters $L = 5$, $r = 1$, and $r_a = 3$ and we consider low ($\eta = 0.005$), moderate ($\eta = 0.05$), and high ($\eta = 0.5$) noise. From the AP trajectories in the top panels of Fig. 1, we see the degradation of alignment among agents with increasing noise, accompanied by a preference for travel along the horizontal axis of the domain in the low noise condition. The SO trajectories are presented in the bottom panels of Fig. 1, where the source appears as a black dot. Due to the relatively wide region of attraction for the source, agents with low noise align and approach the source together. As the noise increases, the alignment among agents decreases. However, the presence of the source induces spatial order among the agents, which repeatedly move toward it after being deflected by the boundaries.

As a validation of this model's ability to foster emergent behavior among agents, we generate simulation data of $N = 40$ noninteracting agents ($r = 0$) affected by neither axial attraction nor a source. In this case, agents' trajectories are independently generated dependent only on random initial conditions and the noise η. We refer to this parameter selection as "random walkers"(RW). Figure 2 presents truncated trajectories of $N = 40$ random walkers subjected to low, moderate, and high noise. Although the low noise condition results in slightly straighter paths for random walkers, the trajectories for all RW simulations show no emergent behaviors among agents such as are evident from the aligned trajectories in the low and moderate noise simulations in Fig. 1.

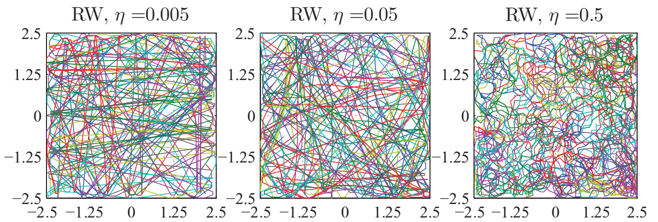

FIG. 2. (Color online) Sample simulation trajectories with RW parameters and $\eta = 0.005$, 0.05, and 0.5. We use $r = 0$ and the same time steps and other protocol parameter values for N, s, L as in Fig. 1. Horizontal and vertical axes refer to the real and imaginary parts of $x(k)$, respectively, for different values of k.

To quantify the order in models of collective behavior, measures based on observed spatiotemporal patterns, such as members of animal groups maintaining aligned traveling directions, selecting proximal positions, or milling about a common center, are typically employed [7].

For example, alignment behavior may be captured as the polarization [32]

$$\mathcal{P}(k) = \frac{1}{N}\left\| \sum_{i=1}^{N} \exp\left[\iota\theta_i(k)\right] \right\|. \tag{5}$$

This quantity equals one when all agents have a common heading and equals zero when the agents can be grouped into pairs with headings that differ by π. In addition, agents' proximity can be measured using the cohesion [47]

$$\mathcal{C}(k) = \frac{1}{N}\sum_{i=1}^{N} \exp\left(\frac{-\|\widetilde{x}_i(k)\|}{L}\right), \tag{6}$$

where $\widetilde{x}_i(k)$ is the relative position of agent i at time k with respect to the center of mass $x_{\text{c.m.}}(k)$. That is,

$$\widetilde{x}_i(k) = x_i(k) - x_{\text{c.m.}}(k), \quad \text{where} \quad x_{\text{c.m.}}(k) = \frac{1}{N}\sum_{i=1}^{N} x_i(k). \tag{7}$$

This quantity equals one when all agents' positions coincide on the group center of mass $x_{\text{c.m.}}(k)$ and goes to zero as the individual distances from $x_{\text{c.m.}}(k)$ increase. We note that L acts as a characteristic length scaling the decay of the cohesion as agents move apart from each other.

For the simulations whose truncated trajectories are shown in Fig. 1, we calculate the polarization and cohesion to assess order in these systems via traditional measures. Statistics of these measures are presented in Table I. Figure 3 presents the polarization for truncated simulations using AP and SO parameters and low, moderate, and high noise. The low and moderate noise conditions maintain polarizations near one for both parameter sets except for occasional deviations which correspond to impacts of the boundary. As the noise reaches its largest value, the polarization drastically decreases as

TABLE I. Measures of simulation complexity using traditional and proposed methods. Simulations are 10 000 time step trials, which are shown truncated in Figs. 1 and 2 using $N = 40$, $s = 0.25$, and $L = 5$. Polarization and cohesion are presented as mean ± one standard deviation. Manifold dimension refers to the embedding manifold identified by the ISOMAP algorithm and the residual variance corresponding to the first dimension is given in the last column.

Parameters	η	$\mathcal{P}$	$\mathcal{C}$	Manifold dimension	Residual dimension = 1
AP	0.005	0.98 ± 0.10	0.97 ± 0.01	1	0.0193
AP	0.05	0.96 ± 0.18	0.84 ± 0.03	6	0.3833
AP	0.5	0.40 ± 0.15	0.73 ± 0.04	>10	0.7216
SO	0.005	0.99 ± 0.07	0.99 ± 0.00	1	0.0143
SO	0.05	0.97 ± 0.09	0.97 ± 0.01	2	0.4906
SO	0.5	0.41 ± 0.15	0.73 ± 0.04	>10	0.7474
RW	0.005	0.13 ± 0.07	0.69 ± 0.02	>10	0.9932
RW	0.05	0.13 ± 0.07	0.69 ± 0.02	>10	0.9963
RW	0.5	0.14 ± 0.07	0.69 ± 0.02	>10	0.9949

expected. Similarly, the cohesion of simulations with different parameter sets offers insight into varying complexity in the system; see Fig. 4. Specifically, simulations with AP and SO parameters show uniformly high cohesion in both low and moderate noise conditions. The high noise condition instead has markedly smaller values of cohesion, as is verified by the statistics in Table I.

The emergence of collective behavior in simulations with AP and SO parameters can be viewed in contrast with the lack of order in simulations with RW parameters; see Fig. 5. As measured by the polarization and cohesion, the random walker simulations show no variations in complexity for different values of η.

III. THE ISOMAP ALGORITHM

The ISOMAP algorithm is applied to a data set comprising an array of n d-dimensional data points with the goal of embedding them on a manifold, assessing the dimensionality of such manifold, and perhaps finding its dimension to be less than d. Specifically, for a data set $\mathcal{Z} = \{z_i\}_{i=1}^n \subset \mathbb{R}^d$, we construct a corresponding data set, $\mathcal{Y} = \{y_i\}_{i=1}^n \subset \mathbb{R}^{\hat{d}}$, appropriately embedded within an invariant manifold, and assess if $\hat{d} \ll d$. In the following, we describe $\mathcal{Z}$ in the variables of the ambient space $\mathbb{R}^d$ and $\mathcal{Y}$ in the intrinsic variables of the manifold, which is locally homeomorphic to $\mathbb{R}^{\hat{d}}$. We embed a point $z \in \mathbb{R}^d$ into intrinsic variables y of a $\hat{d}$-dimensional manifold by representing the manifold in terms of a parametrization,

$$\Phi : \mathcal{Y} \to \mathcal{Z}, \tag{8}$$

where

$$z = \Phi(y) = [\phi_1(y_1, y_2, \ldots, y_{\hat{d}}), \ldots, \phi_d(y_1, y_2, \ldots, y_{\hat{d}})] \tag{9}$$

and $y_i, i = 1, \ldots, \hat{d}$, are intrinsic variables that can be described as directions to any point on the manifold relative to a base point on the manifold.

The ISOMAP algorithm is a manifold learning algorithm that builds classical multidimensional scaling method (MDS) [48] by using approximations of geodesic distances. Rather than directly applying MDS to the ambient Euclidean space, ISOMAP uses shortest paths along a discrete graph approximation of the manifold. There are several steps that are needed to develop the ISOMAP embedding, that is, to represent the parameters $\mathcal{Y}$ in Eq. (9); see the tutorial in [44]. For convenience, we concatenate such parameters into a matrix $Y \in \mathbb{R}^{n \times \hat{d}}$ below.

(i) *Build a neighbors graph to approximate the embedding manifold.* A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of the set of vertices

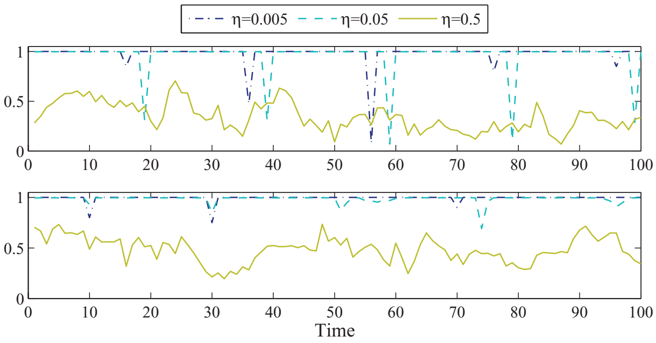

FIG. 3. (Color online) Polarization of simulation trajectories with AP (top) and SO (bottom) parameters and $\eta = 0.005$, 0.05, and 0.5. We display the first 100 time steps of the 10 000 time step simulations in Fig. 1.

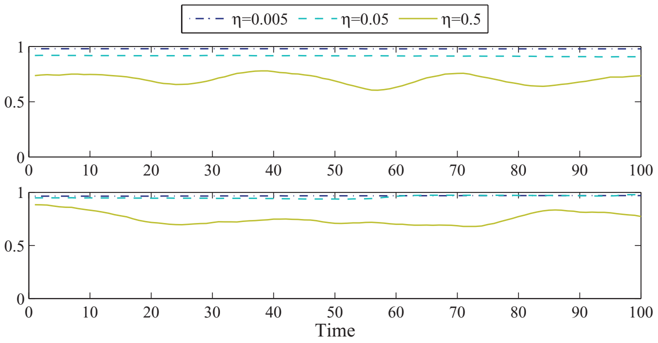

FIG. 4. (Color online) Cohesion of simulation trajectories with AP (top) and SO (bottom) parameters and $\eta = 0.005$, 0.05, and 0.5. We display the first 100 time steps of the 10 000 time step simulations in Fig. 1.

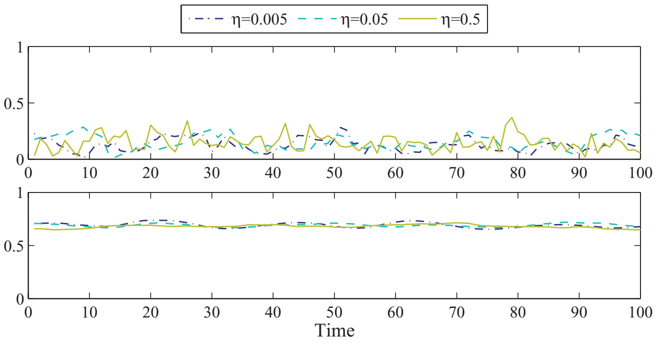

FIG. 5. (Color online) Polarization (top) and cohesion (bottom) of simulation trajectories with RW parameters and $\eta = 0.005$, 0.05, and 0.5. We display the first 100 time steps of the 10 000 time step simulations in Fig. 2.

$\mathcal{V} = \{v_i\}_{i=1}^n$ which we assign to match the data points, $\mathcal{Z} = \{z_i\}_{i=1}^n$ and the set of edges $\mathcal{E}$ that has elements which are unordered pairs of vertices present in the graph. We assign edges to connect vertices which are either ϵ neighbors or ν-nearest neighbors. To build a ν-nearest-neighbors graph, we construct the graph consisting of edges $\{v_i,v_j\}$ corresponding to the ν-closest data points z_j's to z_i, for each i, with respect to the Euclidean distance in the ambient space, which we denote $d_Z(z_i,z_j)$. We let $M_n \in \mathbb{R}^{n\times n}$ be a matrix encoding the weighted graph of intrinsic manifold distances corresponding to the graph $\mathcal{G}$, whose ijth entry is denoted $M_n(i,j)$. For each edge $\{v_i,v_j\} \in \mathcal{E}$, we define the distances $M_n(i,j) \approx d_Z(z_i,z_j)$, and for all $\{v_i,v_j\} \notin \mathcal{E}$, we associate $M_n(i,j) = \infty$ to prevent jumps between branches of the underlying manifold.

(ii) *Compute geodesics of the graph to approximate geodesics of the manifold.* There are popular methods to compute shortest paths of the graph, including Floyd's algorithm for small to medium sized data sets [49] or Dijkstra's algorithms for small to large data sets [50]. Using M_n, we compute an approximate geodesic distance matrix $D_M \in \mathbb{R}^{n\times n}$, whose ijth element consists of the shortest weighted path length between each v_i to v_j, thus approximating manifold geodesic distances.

(iii) *Approximate manifold distance by ν-nearest-neighbor distance.* The distance matrix D_M from the previous step is taken to approximate the true geodesic distances of the manifold between z_i and z_j by the distance between v_i and v_j. This approximation improves as data density increases. If ν is chosen too large or data density is too low, then some neighbors may reside on separate branches of the manifold. In such cases, the approximation is poor due to "illegal" shortcuts and a poor representation of the manifold.

(iv) *Perform an MDS on D_M.* MDS requires only the matrix D_M in manifold distances as input, which is computed from the input data $\mathcal{Z}$ to form projective variables $\mathcal{Y}$ in the intrinsic variables.

For completeness, we review the classical MDS algorithm presented in [48]. Given D_M, which approximates in manifold geodesic distances for our purposes, the goal is to form a matrix of projected d-dimensional data Y to minimize the residual error defined

$$E = \|\tau(D_M) - \tau(D_Y)\|_{L^2}. \tag{10}$$

Here, for $A \in \mathbb{R}^{n\times n}$, $\|A\|_{L^2} \equiv \sqrt{\sum_{i=1}^n \sum_{j=1}^n A(i,j)^2}$ is the Holder matrix two-norm, see for example [51], $\tau(A)$ is a matrix-valued centered distance function, and the ijth entry of the matrix D_Y describes the geodesic distance between y_i and y_j in the projective space of intrinsic variables. In other words, the matrix Y is the result of an optimization problem. The central distance function when applied to D_M is defined by (see [48])

$$\tau(D_M) = -\tfrac{1}{2} H D_M^2 H, \tag{11}$$

where H is a centering matrix

$$H = I_n - \frac{1}{n} 1_n 1_n^{\mathrm{T}}, \tag{12}$$

1_n is the n-dimensional column vector of all ones, I_n is the n-dimensional identity matrix, and superscript T denotes matrix transposition. Upon replacing the definition of central distance function in Eq. (11) into the error in Eq. (10), we find that

$$\begin{aligned} &\min_{Y\in\mathbb{R}^{n\times d}} \|\tau(D_M) - \tau(D_Y)\|_{L^2} \\ &= \min_{Y\in\mathbb{R}^{n\times d}} \big\| -\tfrac{1}{2} H\big(D_M^2 - D_Y^2\big)H\big\|_{L^2} \\ &= \min_{Y\in\mathbb{R}^{n\times d}} \|\tilde{Z}^{\mathrm{T}}\tilde{Z} - Y^{\mathrm{T}}Y\|_{L^2}^2 \\ &= \min_{Y\in\mathbb{R}^{n\times d}} [\mathrm{trace}(\tilde{Z}\tilde{Z}^{\mathrm{T}} - YY^{\mathrm{T}})]^2. \end{aligned} \tag{13}$$

The latter equalities follow from the theorem in [48] that yields that, for any selection of the matrix D_M, there exist a matrix $\tilde{Z} \in \mathbb{R}^{n\times d}$

$$\tau(D_M) = \tilde{Z}^{\mathrm{T}}\tilde{Z}, \tag{14}$$

and likewise for $\tau(D_Y)$. The matrix $\tilde{Z}$ can be understood as centered in such a way that pairwise Euclidean distances are D_M.

A key advantage of the MDS algorithm over the more common proper orthogonal decomposition algorithm is that all matrix manipulations to compute an output Y require only the centered distance matrix $\tau(D_M)$, which represents geodesic distances on the manifold. Therefore, $\tilde{Z}$ is allowed to be in the manifold appropriate to the geodesic distances D_M and $\tilde{Z}$ is thus distinguished from the original input data Z.

Since $\tau(D_M)$ is symmetric and positive semidefinite, the computation of MDS uses the spectral decomposition,

$$\tau(D_M) = V\Sigma V^{\mathrm{T}}, \tag{15}$$

where $\Sigma \in \mathbb{R}^{n\times n}$ is the diagonal matrix composed of the eigenvalues of D_M from the set $\{\lambda_i\}_{i=1}^n$, and $V \in \mathbb{R}^{n\times n}$ is the orthogonal matrix whose columns are eigenvectors and

$$V^{\mathrm{T}}V = I_n \quad \text{and} \quad \tau(D_M)V = \Sigma V. \tag{16}$$

Comparing representations for $\tau(D_M)$, Eqs. (11) and (14), to the spectral decomposition Eq. (15) yields

$$\tilde{Z} = \Sigma^{\frac{1}{2}} V^{\mathrm{T}}, \tag{17}$$

where the square matrix of non-negative eigenvalues has a simple square root $\Sigma = \mathrm{diag}(\sqrt{\lambda_i})$. The MDS solution is then

$$Y \equiv Y_{\mathrm{MDS}} = \Sigma_p^{\frac{1}{2}} V_p^{\mathrm{T}}, \tag{18}$$

where $\Sigma_p^{\frac{1}{2}}$ and V_p use the top p (significant) eigenvalues and eigenvectors of $\tau(D_M)$.

MDS forms the rank-p projection that optimizes the dissimilarity in terms of the intrapoint distances, similarly to principle component analysis (PCA). Specifically, the variables of the corresponding projections relate according to

$$Y_{\text{PCA}} = \Sigma_{\text{PCA}}^{\frac{1}{2}} Y_{\text{MDS}}, \tag{19}$$

where $\Sigma_{\text{MDS}} = \Sigma_{\text{PCA}} = \Sigma_p$ used in Eq. (18). Also, there is a relationship of the basis vectors,

$$V_{\text{PCA}} = \tilde{Z} V_{\text{MDS}}, \tag{20}$$

where similarly $V_{\text{MDS}} = V_p$ from Eq. (18). The two algorithms yield essentially the same result when the distance matrix is the Euclidean distance, but since we take D_M to be discretely approximated in manifold distance in ISOMAP, the results are different, as are the steps of computation. The most important difference in the algorithmic steps between MDS and PCA is that MDS does not explicitly use $\mathcal{Z}$ in its computations. Since variables to be found are in some unknown nonlinear manifold, this is a prudent dependency to avoid.

The outputs of MDS, and therefore the ISOMAP algorithm, are an embedding manifold for the data set and residual variances quantifying the proportion of data points which do not lie on such manifold. The density of data points in the embedding manifold illustrates whether the size of the data set is sufficient to ascertain low dimensionality with respect to ν and the residual variances are the percentages of data points which are not captured by an embedding manifold of a given dimension. To retrospectively assess whether the data set is sufficiently dense to perform the ISOMAP algorithm with ν-nearest neighbors, we examine the embedding manifolds and ensure that no degenerate low-dimensional structures resulting from paucity of data or improper selection of ν exist. The overall procedure is synoptically illustrated in Fig. 6.

To identify the dimensionality of an embedding manifold which well approximates a data set, we seek an "elbow" in the curve of residual variances, after which residual variances do not significantly decrease. The dimension corresponding to such an elbow gives the reduced dimensionality of the data set uncovered by the ISOMAP algorithm.

IV. RESULTS

For simulations with AP, SO, and RW parameters, we acquire the two-dimensional positions of the agents and implement the ISOMAP algorithm on their distribution in a discretized spatial domain, where the domain $[-L/2,L/2] \times [-L/2,L/2]$ is partitioned into a square two-dimensional grid of 50×50 cells. Analogous to an image on a screen which encompasses multiple pixels, we enlarge the position of each agent to include 5 cells in both directions so that each agent resides in an 5×5 moving square. This strategy is employed to discretize the distances between data points and thus restrict the number of vertices in the graph calculated by the ISOMAP algorithm. At every time step, the entries of the 1×2500 position distribution vector report the number of agents residing in a given cell.

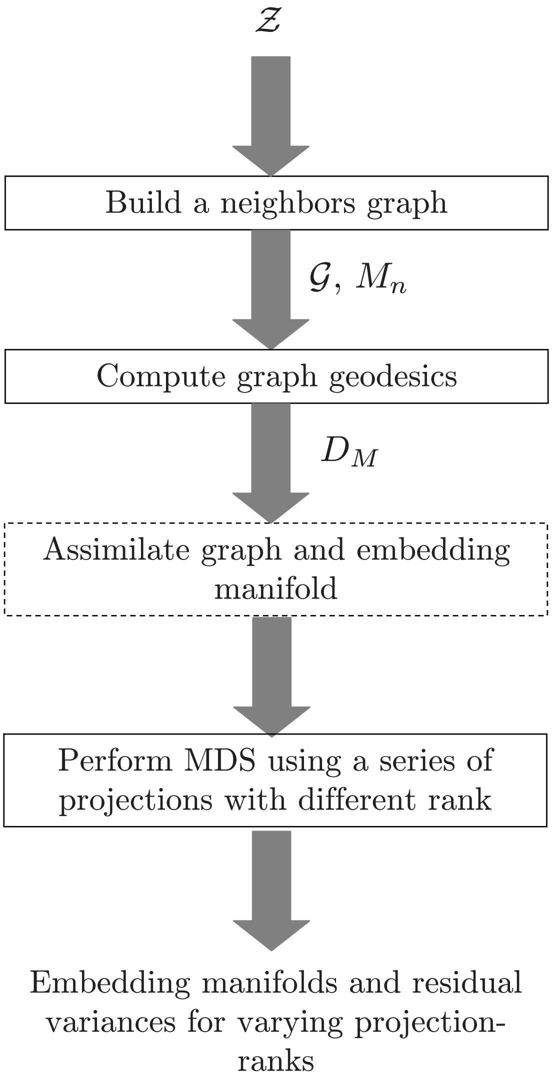
$\mathcal{Z}$
Build a neighbors graph
$\mathcal{G}$, M_n
Compute graph geodesics
D_M
Assimilate graph and embedding manifold
Perform MDS using a series of projections with different rank
Embedding manifolds and residual variances for varying projection-ranks

FIG. 6. Illustration of the main steps for implementation of the ISOMAP algorithm.

The ISOMAP algorithm from [46] is executed on the time trace of this vector by using the number of nearest neighbors $\nu = 11$ and the number of time steps $\tau = 10000$, corresponding to the whole simulation duration. The algorithm is implemented on the raw data without requiring knowledge about the number of agents, the interaction rules, the domain size, and the boundary conditions. The value of ν is selected based on the expectation of a low-dimensional embedding manifold and the parameter value of τ is large enough so that artificial low-dimensional manifolds based on paucity of data are excluded. For nearly noiseless particles, that is $\eta \sim 0$, many nearly identical data points may recur within the data set which confounds nearest-neighbor selection in the ISOMAP algorithm; we select the low noise $\eta = 0.005$ to prevent this scenario.

The two-dimensional embedding manifolds generated by the ISOMAP algorithm from model data using AP and SO parameters and the three considered values for η are given in Fig. 7. From the density of these two-dimensional manifolds, we conclude that the size of the data set is sufficient to perform the ISOMAP algorithm. We notice that the manifolds, scaled between -1 and 1 in each case, cover

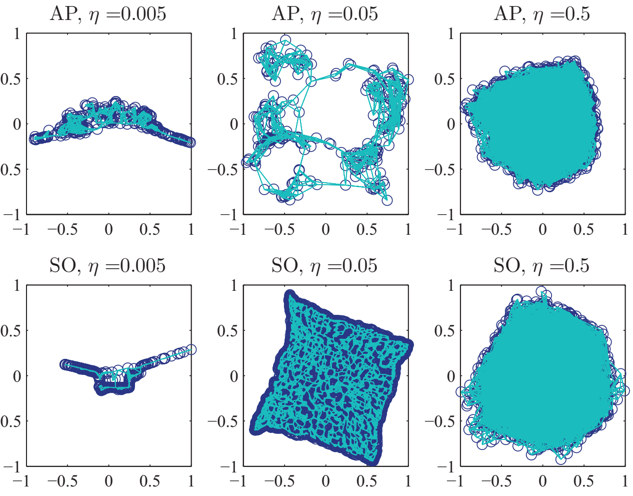

FIG. 7. (Color online) Downsampled two-dimensional embedding manifolds for model with AP and SO parameters and $\eta = 0.005$, 0.05, and 0.5. Simulations are those which are shown truncated in Fig. 1, here sampled at every other time step, and thus use the same model parameters.

a higher proportion of the domain as noise in the model increases. Continuing this trend, the embedding manifolds from simulations using RW parameters cover the domain comparably to those corresponding to AP and SO simulations with high noise; see Fig. 8.

The process of deciding on an appropriate embedding dimension for the algorithm involves testing progressively higher-dimensional embedding manifolds. Figure 9 presents the residual variances varying with increasing embedding manifold dimensionality and scaled according the residual variance at dimensions one; numerical results from this analysis are reported in Table I. We select a decrease of the residual variance to less than 0.05 as the location of the elbow. Thus the dimensionality of the AP embedding manifolds is approximately equal to 1 for low noise, 6 for moderate noise, and greater than 10 for high noise. For the SO case, the residual variances yield embedding manifolds with dimensionality of 1 for low and 2 for moderate noise values and greater than 10 for high noise. In addition, as noise increases, both of these parameter cases exhibit a dramatic increase in the residual variance corresponding to dimension one. In contrast, simulations with RW parameters exhibit a lack of elbows in residual variance curves and large magnitude of the residual variances corresponding to dimension one for all values of η.

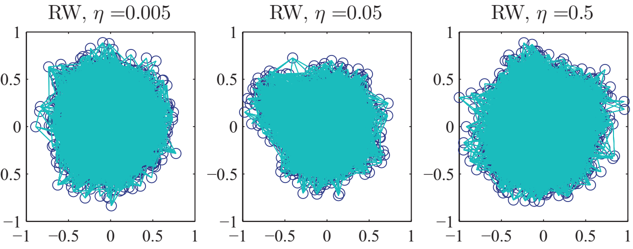

FIG. 8. (Color online) Downsampled two-dimensional embedding manifolds for model with RW parameters and $\eta = 0.005$, 0.05, and 0.5. Simulations are those which are shown truncated in Fig. 2, here sampled at every other time step, and thus use the same model parameters.

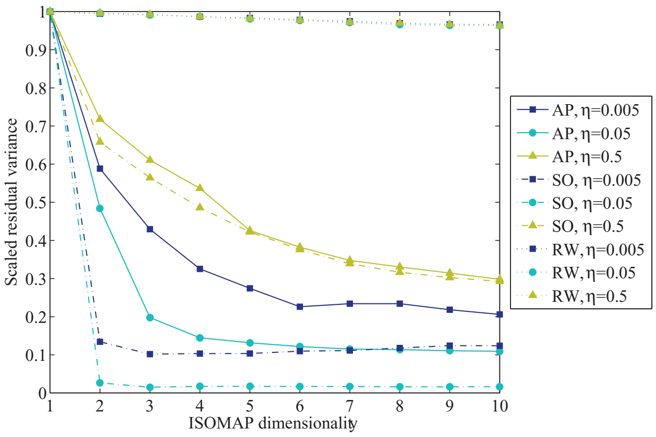

FIG. 9. (Color online) Scaled residual variances for model with AP, SO, and RW parameters and $\eta = 0.005$, 0.05, and 0.5. Residual variances are scaled with respect to those corresponding to dimension one.

V. DISCUSSION

Based on the presented results, the ISOMAP algorithm is able to identify complexity inherent in multiagent systems exhibiting collective behavior. Using traditional measures of order based on polarization and cohesion of agents, we corroborate the intuitive notion that increasing random noise in a system increases its complexity. This trend is evident in simulations using both AP and SO parameters, which suggests that the noise acts uniformly on the interacting-particle model independent of other protocol parameters. In turn, the ISOMAP algorithm differentiates these cases by reporting higher-dimensional embedding manifolds for data sets corresponding to simulations with higher noise. We comment that the order parameters of cohesion and polarization are informed by observations of the group behavior, that is, from *a priori* knowledge of the nature of the emergent behavior. In contrast, the approach proposed in this work, based on machine learning tools, prescinds from any knowledge of the underlying phenomena by using only raw data as input.

Moreover, the ISOMAP algorithm easily distinguishes between simulations with interacting and noninteracting agents, as can be seen comparing simulations with RW parameters to AP and SO parameters. When the agents are random walkers, no emergent behaviors result from their interaction, as evidenced by the uniformly low values for polarization and cohesion. Indeed, this disorder is corroborated by the high residual variances corresponding to dimension one and embedding manifold dimensionalities given by the ISOMAP algorithm in these cases.

The ISOMAP algorithm suggests higher-dimensional embedding manifolds for data pertaining to AP rather than SO parameters. This hints at different levels of complexity in these collective behaviors as a result of the interplay of noisy consensus dynamics, exogenous stimuli, and boundary effects. Notably, such differences in dimensionality cannot be discerned from simple inspection of the truncated trajectories

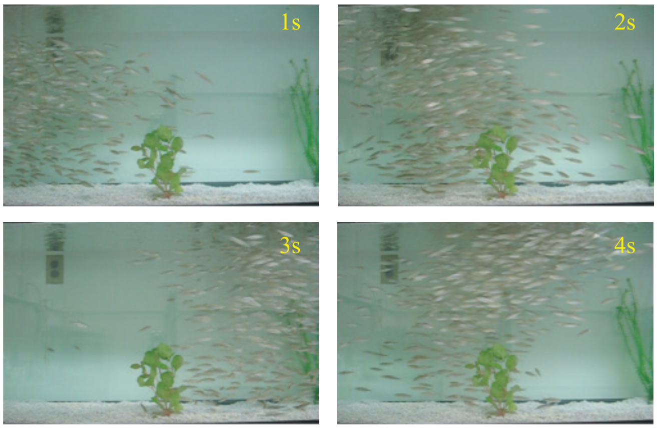

FIG. 10. (Color online) Images of fish schooling experiment. Sample frames of a 52 s movie of a fish school making a circuit of the aquarium.

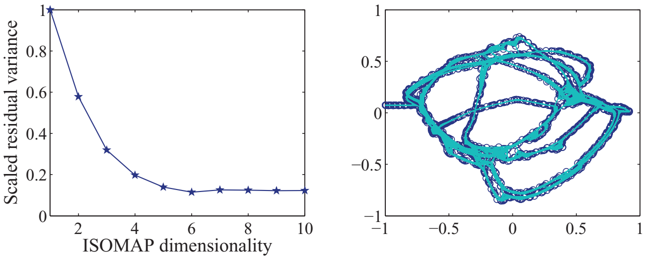

FIG. 11. (Color online) Data from fish schooling experiment. Scaled residual variances (left) and two-dimensional embedding manifold (right) for movie frame data. The residual variances are scaled with respect to that corresponding to dimension one, which equals 0.5894. The movie was filmed at 30 frames/s for 1580 frames and frame size is 480 × 640 pixels.

in Fig. 1 nor from the mean values of polarization and cohesion in Table I. However, such qualitative differences are uncovered by a systematic application of the dimensionality reduction algorithm.

We note that the reflective boundary conditions, in contrast to the periodic boundaries in [32], hinder the persistence of agents' alignment and thus may act as an additional detriment to low-dimensional structures. Hallmarks of this boundary condition are the periodic dips in polarization in Fig. 3, which correspond to the group of agents impacting the boundary and reorienting their headings identically and almost simultaneously. Such coordinated reorientations are absent in the high noise conditions since agents fail to move as a cohesive unit and thus do not simultaneously encounter the boundary. If reflective boundary conditions were removed in favor of periodic ones, intuition would suggest that the agents' motion would be well described by a one-dimensional manifold for a larger range of noise.

To demonstrate the potential impact of this definition of complexity with experimental animal groups, we execute the ISOMAP algorithm on image data of a school of live minnows swimming in synchrony. Snapshots of the fish school used for this analysis are presented in Fig. 10. The movie in [52] records a several-hundred-member fish school moving across the tank approximately five times in 50 s. We apply the ISOMAP algorithm to 480 × 640 pixel values in the range 1 to 256 for the gray scale image at each of the 1580 frames using $\nu = 11$ and Fig. 11 displays the results. The resulting two-dimensional embedding manifold appears dense, thus suggesting enough data to validate the residual variances computed by the ISOMAP algorithm. In fact, the scaled residual variances in the left panel in Fig. 11 show remarkable similarity with those from the moderate noise model with AP parameters in Fig. 9, whose agents execute similar maneuvers to the live fish school. This finding confirms the proposed data-driven working definition of collective behavior as the manifestation of a low-dimensional manifold underlying what an untrained individual would classify as fish schooling. By using the hard threshold of 0.05 for the drop in the residual variance, we find that the embedding manifold dimension is greater than 10. Nevertheless, the considerable reduction to approximately 0.07 for a dimensionality of 6 and the finite size of the data set suggests the existence of a lower-dimensional embedding.

In conclusion, we have introduced a working definition of collective behavior based on dimensionality reduction of large-scale data sets. This definition has been tested against traditional measures of collective behavior on numerical simulations whose parameters are selected to represent six different modalities of interaction. Using the ISOMAP algorithm, we have identified fundamental differences in the embedding manifolds of these data sets compared to one another and to the trivial model in which all interactions among agents are absent, which supports the proposed definition of collective behavior. This method has then been tested on image data from a live school, where we have found low-dimensional structures similar to those observed in the numerical study. Future work will include applying the ISOMAP algorithm to larger experimental data sets on animal groups exhibiting a variety of different behaviors identified by a human observer. Based on the preservation of geodesics by the ISOMAP algorithm and this work, we expect such analysis to uncover the few fundamental parameters which dictate such typical behaviors.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. CMMI-1129820, Grant No. CMMI-0745753, and GK-12 Fellows Grant No. DGE-0741714. This research was also supported in part by New York University SEED funding.

[1] B. L. Partridge, Sci. Am. **246**, 114 (1982).

[2] H. P. Zhang, A. Be'er, R. S. Smith, E.-L. Florin, and H. L. Swinney, Europhys. Lett. **87**, 48011 (2009).

[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, Proc. Natl. Acad. Sci. USA **105**, 1232 (2008).

[4] D. J. T. Sumpter, *Collective Animal Behavior* (Princeton University Press, Princeton, NJ, 2009).
[5] D. Weihs, Nature (London) **241**, 290 (1973).
[6] J. Krause and G. Ruxton, *Living in Groups* (Oxford University Press, New York, 2002).
[7] T. J. Pitcher and J. K. Parrish, *The Behaviour of Teleost Fishes* (Chapman and Hall, London, 1993), Chap. Functions of Shoaling Behaviour in Teleosts, pp. 363–440.
[8] J. Krause, J.-G. J. Godin, and D. Brown, J. Fish Biol. **49**, 221 (2005).
[9] J. Krause, A. J. Ward, A. L. Jackson, G. D. Ruxton, R. James, and S. Currie, J. Fish Biol. **67**, 866 (2005).
[10] C. Leblond and S. G. Reebs, Behaviour **143**, 1263 (2006).
[11] R. F. Lachlan, L. Crooks, and K. N. Laland, Animal Behaviour **56**, 181 (1998).
[12] S. G. Reebs, Animal Behaviour **59**, 403 (2000).
[13] S. G. Reebs, Behaviour **138**, 797 (2001).
[14] B. L. Partridge, Animal Behaviour **28**, 68 (1980).
[15] D. J. T. Sumpter, J. Krause, R. James, I. D. Couzin, and A. J. W. Ward, Curr. Biol. **18**, 1773 (2008).
[16] R. W. Tegeder and J. Krause, Philos. Trans. R. Soc. London: Biol. Sci. **350**, 381 (1995).
[17] S. Torisawa, T. Takagi, H. Fukuda, Y. Ishibashi, Y. Sawada, T. Okada, S. Miyashita, K. Suzuki, and T. Yamane, J. Fish Biol. **71**, 411 (2007).
[18] I. Aoki, Bull. Jpn. Soc. Sci. Fish. **48**, 1081 (1982).
[19] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin, Nature (London) **433**, 513 (2005).
[20] U. Erdmann, W. Ebeling, and A. S. Mikhailov, Phys. Rev. E **71**, 051904 (2005).
[21] W. Li, IEEE Trans. Syst., Man Cybernet.- Part B: Cybernet. **38**, 1084 (2008).
[22] R. P. Mann, PLoS ONE **6**, e22827 (2011).
[23] H.-S. Niwa, J. Theor. Biol. **181**, 47 (1996).
[24] A. Perez-Escudero and G. G. de Polavieja, PLoS Computat. Biol. **7**, e1002282 (2011).
[25] M. T. Rashid, M. Frasca, A. A. Ali, R. S. Ali, L. Fortuna, and M. G. Xibilia, Nonlin. Dynam., doi: 10.1007/s11071-011-0237-6.
[26] M. Zheng, Y. Kashimori, O. Hoshino, K. Fujita, and T. Kambara, J. Theor. Biol. **235**, 153 (2005).
[27] W. Ren and R. W. Beard, *Distributed Consensus in Multi-vehicle Cooperative Control* (Springer-Verlag, London, 2008).
[28] E. Erkip, A. Sendonaris, A. Stefanov, and B. Aazhang, *Advances in Network Information Theory* (American Mathematical Society, Providence, 2004), Chap. Cooperative Communication in Wireless Systems, pp. 303–320.
[29] S. Massoud Amin and B. F. Wollenberg, IEEE Power Energy Mag. **3**, 34 (2005).
[30] Project PigeonWatch, http://www.birds.cornell.edu/, pigeonwatch.
[31] G. J. Stephens, B. Johnson-Kerner, W. Bialek, and W. S. Ryu, PLoS Computat. Biol. **4**, e1000028 (2008).
[32] T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet, Phys. Rev. Lett. **75**, 1226 (1995).
[33] M. R. D'Orsogna, Y. L. Chuang, A. L. Bertozzi, and L. S. Chayes, Phys. Rev. Lett. **96**, 104302 (2006).
[34] F. Ginelli and H. Chate, Phys. Rev. Lett. **105**, 168103 (2010).
[35] A. Kolpas, J. Moehlis, and I. G. Kevrekidis, Proc. Natl. Acad. Sci. USA **104**, 5931 (2007).
[36] D. Morgan and I. Schwartz, Phys. Lett. A **340**, 121 (2005).
[37] J. K. Parrish, S. V. Viscido, and D. Grunbaum, Biol. Bull. **202**, 296 (2002).
[38] M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum http://pages.pomona.edu/∼vds04747/public/papers/BdSLT.pdf (unpublished).
[39] S. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[40] J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290**, 2319 (2000).
[41] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, in *Proceedings of the International Conference on Artificial Neural Networks*, edited by F. Fogelman and P. Gallinari (EC2 & Cie, Paris, France, 1995), pp. 53–60.
[42] B. Schlkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)* (MIT Press, Cambridge, MA, 2001).
[43] Y. Gong and W. Xu, *Machine Learning for Multimedia Content Analysis* (Springer-Verlag, New York, 2007).
[44] E. Bollt, Int. J. Bifurcat. Chaos **17**, 1199 (2007).
[45] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, MA, 2004).
[46] A Global Geometric Framework for Nonlinear Dimensionality Reduction, http://web.mit.edu/cocosci/isomap/isomap.html.
[47] M. Aureli and M. Porfiri, Europhys. Lett. **92**, 40004 (2010).
[48] T. F. Cox and M. A. Cox, *Multidimensional Scaling* (Chapman and Hall, London, 1994).
[49] R. W. Floyd, Commun. ACM **5**, 345 (1962).
[50] E. W. Dijkstra, Numer. Math. **1**, 269 (1959).
[51] G. G. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (Johns Hopkins, Baltimore, MD, 1996).
[52] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.85.041907 for a representative video clip of the fish school.