Vicsek model by time-interlaced compression: A dynamical computable information density

A. Cavagna,[1,2] P. M. Chaikin,[3] D. Levine,[4] S. Martiniani,[5] A. Puglisi,[1,2] and M. Viale[1,2]

[1] *Dipartimento di Fisica, Università La Sapienza, 00185 Rome, Italy*
[2] *Istituto dei Sistemi Complessi, Consiglio Nazionale delle Ricerche, 00185 Rome, Italy*
[3] *Center for Soft Matter Research, Department of Physics, New York University, New York, New York 10003, USA*
[4] *Department of Physics, Technion-IIT, 32000 Haifa, Israel*
[5] *Department of Chemical Engineering & Materials Science, University of Minnesota, Minneapolis, Minnesota 55455, USA*

(Received 24 August 2020; accepted 13 May 2021; published 25 June 2021)

Collective behavior, both in real biological systems and in theoretical models, often displays a rich combination of different kinds of order. A clear-cut and unique definition of "phase" based on the standard concept of the order parameter may therefore be complicated, and made even trickier by the lack of thermodynamic equilibrium. Compression-based entropies have been proved useful in recent years in describing the different phases of out-of-equilibrium systems. Here, we investigate the performance of a compression-based entropy, namely, the computable information density, within the Vicsek model of collective motion. Our measure is defined through a coarse graining of the particle positions, in which the key role of velocities in the model only enters indirectly through the velocity-density coupling. We discover that such entropy is a valid tool in distinguishing the various noise regimes, including the crossover between an aligned and misaligned phase of the velocities, despite the fact that velocities are not explicitly used. Furthermore, we unveil the role of the time coordinate, through an encoding recipe, where space and time localities are both preserved on the same ground, and find that it enhances the signal, which may be particularly significant when working with partial and/or corrupted data, as is often the case in real biological experiments.

DOI: 10.1103/PhysRevE.103.062141

I. INTRODUCTION

Statistical physics and information theory have a long history of cross-fertilization [1], with both disciplines based on a key concept: a quantitative statistical measure of order [2,3]. For physical systems both in and out of equilibrium, such a measure is the starting point for a general theory of phase transitions, and it is a prerequisite for the study of response to external perturbations [4]. The statistical notion of entropy—the fundamental link between thermodynamics and equilibrium statistical mechanics—is the main inspiration for the central concept of information theory, the Shannon entropy [5]. More recently, information theoretic ideas have proven useful in the study of physical many-body systems, for example, to obtain good estimates for critical temperatures in equilibrium spin systems [6,7], to describe entropy production in the context of stochastic thermodynamics [8,9], and to obtain accurate estimates of the entropy in both equilibrium and nonequilibrium systems [10,11].

In this paper, we seek a method to measure order in some nonequilibrium flocking models, with an eye to eventually analyzing observational data on living systems. In particular, we propose an analysis which treats ordering in space and time on an equal footing. We argue that by analyzing the temporal development of spatial patterns together, we can obtain information in a way that is robust enough to survive the inevitable noise present in collections of living objects.

Our approach is based on a recently proposed information-based measure of order for out-of-equilibrium systems [10]. This proposal, called *computable information density* (CID), relates to the compression rate measured by lossless compression algorithms [12,13]. CID was shown to give clear signatures of important transitions in the systems studied, which included several absorbing state models [14] as well as an active matter model, repulsive active Brownian particles (ABPs), where motility-induced phase separation appears at large enough concentrations [15]. We note that these models lack first-principles Hamiltonians, which makes this quantification of order even more compelling.

In particular, we compress a sequence s of L bits by using a universal lossless compression algorithm (such as one of the Lempel-Ziv algorithms) [16,17] and define the CID of s as

$$\mathrm{CID}(s) \equiv \frac{\mathcal{L}(s)}{L}, \tag{1}$$

where $\mathcal{L}(s)$ is the total binary code length of the compressed sequence. We note that although we are using the term "sequence," we are not restricted to one-dimensional (1D) strings: Microstates in any dimension may be compressed by appropriate procedures, as discussed later. For equilibrium systems, the CID gives a good approximation to the thermodynamic entropy [10].

Here, we investigate the effectiveness of CID in flocking models, a class of active matter known to exhibit a different kind of order from ABPs. We introduce a "space-time" algorithm based on CID which is sensitive enough to yield insight even when it only uses strongly coarse-grained information about the system, such as a binarized (empty vs occupied) density field. Our main results are as follows:

(i) By compressing a proper coarse-grained representation of the system that only considers occupied vs unoccupied regions of space, we introduce a CID-derived observable (which we indicate with the symbol Q) which is able to characterize the amount of correlations in the system.

(ii) We observe that the behavior of Q correlates with the different nontrivial ordering phenomena in flocking models and that it detects multiple phases in the system.

(iii) We are able to improve the sensitivity of Q by proper incorporation of temporal as well as spatial information in the representation of the system, and in this way we capture information about temporal correlations as well.

(iv) This increase in sensitivity allows us to measure important features even when our data are substantially corrupted, a condition typically encountered in data collected from actual living systems.

In Sec. II we introduce the flocking model used to test our approach, the two-dimensional Vicsek model, following which we explain the encoding method we use to translate its dynamical configurations into a binary string, and we then define our order measure. In Sec. III we present results obtained from numerical simulations. We draw our conclusions in Sec. IV.

II. ENCODING DYNAMICAL CONFIGURATIONS

A. The Vicsek model

We consider here an archetypical model of collective behavior in biological systems, the Vicsek model (VM) [18], which we study in two dimensions. The VM is an active matter model in which each agent or particle moves at fixed speed and updates its velocity orientation by imitating the average velocity of its nearest neighbors. The first is a simple self-propulsion mechanism, while the second is an alignment mechanism inspired by classical ferromagnetic models with continuous symmetry [19,20]. Such features lead to a general continuous theory of flocking [21–24]. It is worth noting that the interaction network (or connectivity matrix) depends on time through the interparticle distances, which in turn depend on the velocities; this closes a feedback loop between positions and velocities that drives the systems out of equilibrium. When we add noise to the update rules, the VM displays a transition between a high-noise disordered phase and a low-noise ordered (or polarized) phase. Close to this transition, nontrivial fluctuations emerge both in velocity and in density, whose properties at finite size depend on the nature of the noise and interaction [25,26].

More specifically, the model consists of a system of N active particles, each moving with fixed speed v_0 in a 2D square box of area $W \times W$ with periodic boundary conditions. At each time step the particles tend to align their direction of motion with that of their neighbors, with some noise added to make the dynamics stochastic. Let $\mathbf{r}_i(t) \equiv (x_i(t), y_i(t))$ and $\mathbf{v}_i(t) \equiv (v_0 \cos\theta_i(t), v_0 \sin\theta_i(t))$ be the position and velocity, respectively, of particle i at some time t and $\theta_i(t)$ be its orientation, and let $\mathcal{N}_i(t)$ be the set of its neighbors in a circle of radius R centered about the particle i (hence this is a *metric* implementation of VM) and $\mathbf{V}_i(t) \equiv \sum_{j\in\mathcal{N}_i(t)} \mathbf{v}_j(t)/|\mathcal{N}_i(t)|$ be the average velocity of the particles in the neighborhood of i. In the original Vicsek implementation, each particle in the system evolves according to the following update rules:

$$\theta_i(t+1) = \Theta[\mathbf{V}_i(t)] + \eta\, \xi_i(t),$$
$$\mathbf{r}_i(t+1) = \mathbf{r}_i(t) + \mathbf{v}_i(t+1),$$

where $\Theta[\mathbf{v}] \equiv \mathrm{Arg}(v_x + iv_y)$ is the angular coordinate of $\mathbf{v}$, $\xi_i(t)$ is a uniformly distributed random variable in $[-\pi, \pi]$, and $\eta \in [0, 1]$ is the noise strength: This is the case of *intrinsic* (also called *scalar*) noise. In this paper we also study another popular noise implementation for VM introduced in Ref. [27]: *extrinsic* (also called *vectorial*) noise. For extrinsic noise, the updating rule for θ_i is a little different:

$$\theta_i(t+1) = \Theta\left[\mathbf{V}_i(t) + \eta \begin{pmatrix} \cos\xi_i(t) \\ \sin\xi_i(t) \end{pmatrix}\right].$$

Both types of noise generate a phase diagram characterized by two critical noise values $\eta_b < \eta_c$ which, finite-size effects apart, depend on the density ρ of the system and the speed v_0 of the particles.

When the noise intensity is large enough, $\eta > \eta_c$, a Vicsek fluid stays in a fully disordered state with a spatially homogeneous density and a thermal-like distribution of particle directions. By reducing the noise to $\eta \simeq \eta_c$ the rotational symmetry is spontaneously broken, and the system exhibits collective motion: A clear polarization transition appears for the velocities. This is accompanied by strong modifications of the density field, similar to microphase separation with traveling band formation in the noise interval $[\eta_b, \eta_c]$. When the noise is further reduced to $\eta < \eta_b$, the bands disappear, replaced by a polarized state whose density is spatially homogeneous but possessing giant fluctuations [28,29]. The presence of microphase separation implies a first-order transition, but there is a difference between the two kinds of noise: At finite size W, sharp transitions for extrinsic noise and smoother transitions for intrinsic noise are usually found [30,31]. Basically, in the case of the intrinsic noise model, the traveling bands are less clear and defined, though they become more and more evident as the size of the system increases, and the transition is said to be weakly first order. The extrinsic noise, on the other hand, exhibits a band structure even at moderate sizes, and the first-order nature of the transition is more evident [27]. In this paper we show how we can probe the phases of VM through a single observable related to the CID of a suitable coding of system configurations and how the behavior of this observable is able to distinguish both the noise implementations. In Fig. 1 we show the polarization and its fluctuations for all the simulated systems of this paper. The peak of the fluctuations localizes the critical value η_c. Locating η_b is usually more complicated,

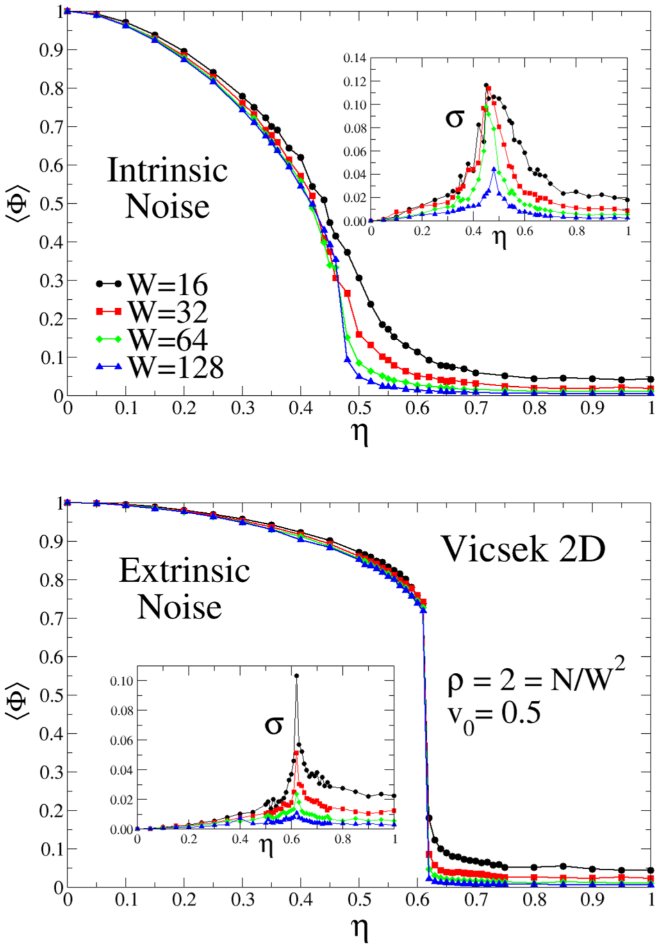

FIG. 1. Mean $\langle\Phi\rangle$ of the order parameter $\Phi = |\sum_i e^{i\theta_i}/N|$ and its fluctuations $\sigma^2 = \langle(\Phi - \langle\Phi\rangle)^2\rangle$ for several system sizes W, as a function of noise strength η, for intrinsic (top) and extrinsic (bottom) cases.

and the local density distribution and/or the hysteresis of the order parameter are typically correlated to it [27,30].

B. Coarse graining

To implement the CID measurements, we first select a space portion of size $l < W$, and we discretize this subset by overlaying a regular square grid of $M = m \times m$ cells of size $b = l/m$. We then observe the system evolution for a time window of T steps. At each time t we assign the symbol "1" to all the cells occupied by at least one particle and the symbol "0" to empty ones: In this way we build a set of $L = M \times T$ bits which stores the evolution of the coarse-grained density field of the system over T steps. With a little abuse of notation, we refer to this set of L bits which encodes the T-step evolution of the system as the "configuration,' and we use C to indicate it and C_{xyt} to indicate its entries [$C_{xyt} = 1$ if there are some particles in the cell at position (x, y) at time t; $C_{xyt} = 0$ otherwise].

We note that we could have considered a richer alphabet, e.g., based on combinations of local density and local (e.g., cell averaged) velocity, but there are good reasons for not doing this. First, in real biological data, we typically have direct access only to the positions of the agents; other degrees of freedom (e.g., velocities) are obtained from the knowledge of the positions. Additionally, since we simultaneously encode $T > 1$ steps, we expect the velocity information to be present implicitly.

C. Scanning: The Z-order curve

Since the typical compression programs operate on one-dimensional strings of characters, we need to scan the three-dimensional array C of L bits in order to produce a one-dimensional sequence. Different scanning procedures exist, but in this paper we employ two procedures based on the so-called *Z-order* or *Morton-order* mapping [32], which is similar to Hilbert scanning. This class of mappings has the advantages of preserving spatial locality in a reasonable fashion and of working in arbitrary dimensions. We discuss the *Z-order* mapping for the 2D case; generalization to higher dimensions is immediate.

Let G be a matrix and G_{xy} be its entries (x and y are integers; in our case, the entries G_{xy} take on the values 0 or 1, but this is not important for the scan). We wish to compose a 1D string $s = s_1s_2s_3\cdots$ which are derived from G. The Z-order algorithm does this as follows:

(1) Write the integer coordinates x and y in binary representation, such that x_i and y_i are the ith digits in the representations.

(2) Interleave the digits of the x and y to form a new binary string $x_1y_1x_2y_2x_3y_3\cdots$ (if the binary string for x or y is shorter than the other, pad it with zeros): This is the binary representation of some integer k.

(3) Set $s_k = G_{xy}$.

The generalization to higher dimensions is straightforward: The binary representations of the lattice site $(x, y, z, \ldots)$ are interleaved in the same fashion as above. In Fig. 2 we show how the Z-order curve works in two dimensions. We note an ambiguity which we will return to later: For a grid of dimension D, there are $D!$ ways to obtain a Z-order curve, one for each permutation of the cell coordinates. In this paper, the configuration C includes a sequence of $T > 1$ different time steps in the evolution of the system. We consider two main ways to scan this three-dimensional matrix of L entries (see Fig. 3):

Serialized time coding (STC). We scan the M cells of each time step according to the 2D Z-order algorithm, and then we concatenate the resultant 1D strings in a sequence of $L = T \times M$ bits according to their time order. In this way, spatial locality is preserved.

Interlaced time coding (ITC). We try to enhance preservation of time locality by scanning the entire space-time set of $M \times T$ cells with the 3D Z-order algorithm and producing a 1D sequence of L bits.

Finally, regardless of how the sequence s was constructed, we compute the CID by the Lempel-Ziv algorithm LZ77 [17] as described in the Appendix, and we study the quantity $Q(s)$ defined by

$$Q(s) = 1 - \mathrm{CID}(s)/\overline{\mathrm{CID}(s_{\mathrm{sh}})},$$

where $\mathrm{CID}(s_{\mathrm{sh}})$ is the CID of a random shuffle s_{sh} of the sequence s and $\overline{\mathrm{CID}(s_{\mathrm{sh}})}$ indicates an average over several

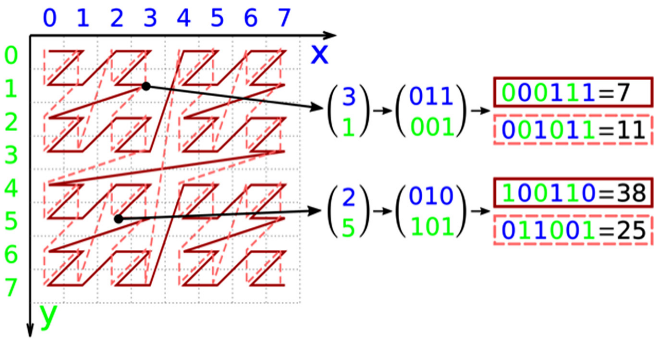

FIG. 2. Procedure to get the Z value, i.e., the coordinate along the Z-order curve which is a one-dimensional spanning of a multidimensional space, preserving locality. Each point has a Z value which is obtained by interleaving the bits of the x coordinate with the bits of the y coordinate. This, in two dimensions, can be done in two different ways: Bits can be interleaved in the order $xyxyxy\cdots$ (x first) or in the order $yxyxyx\cdots$ (y first). The two possibilities are illustrated for two different points, using colors (green and blue) to make appreciable the choice of the interleaving order. One choice gives us the solid curve; the other gives us the dashed curve. In the example, the distance (along the Z curve) between the two points depends on the choice of the order: It is $38 - 7 = 31$ for x-first order and $25 - 11 = 14$ for y-first order.

such shuffled sequences. Q has the characteristic of an order parameter: $0 \leqslant Q \leqslant 1$ and, for asymptotically long strings, $Q \simeq 0$ indicates that the symbols are uncorrelated while $Q > 0$ indicates the presence of some order in the sequence. We note that for STC there are two possible interleavings and for ITC there are $3! = 6$ interleavings. This means that the Z curve does not span the space (or the space-time for the ITC case) in an isotropic way. In order to handle this ambiguity and improve the isotropy, we can average Q over the two (for STC) or six (for ITC) possible curves obtained by different interleavings of coordinates. So, if indicating by $U_S(C)$ and $U_I(C)$ the set of possible strings obtained from the configuration C with STC and ITC schemas, respectively, we define

$$Q_S(C) = \frac{1}{2}\sum_{s\in U_S(C)} Q(s), \quad Q_I(C) = \frac{1}{6}\sum_{s\in U_I(C)} Q(s).$$

It is useful to introduce the following quantities as well:

$$Q^{<}_{S,I}(C) = \min_{s\in U_{S,I}(C)} Q(s), \quad Q^{>}_{S,I}(C) = \max_{s\in U_{S,I}(C)} Q(s).$$

We show below that the ITC schema is more sensitive to the ordering than the STC schema, which we argue will play a useful role in analyzing collective dynamics of living things.

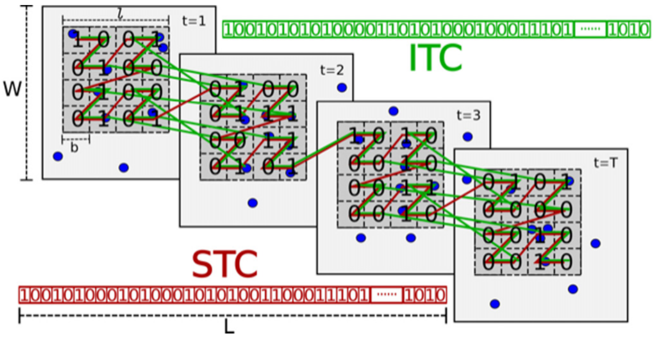

FIG. 3. Comparing STC (red curve) and ITC (green curve) schemes. STC simply places the evolving configurations one behind the other after a two-dimensional Z order spanning over the cells of a single time step. Differently, ITC interlaces bits of configurations at different times following a three-dimensional Z-order curve. Note that the numbers of 1s and 0s are identical and only the positions in the output one-dimensional string are different.

III. RESULTS

In this section we show how the dynamical information arises by an analysis of Q by comparing the ITC and STC schemes on simulations of the 2D Vicsek model. For both intrinsic and extrinsic types of noise, we focus on a typical set of parameters, setting the interaction radius, density, and speed to $R = 1$, $\rho = 2$, and $v_0 = 0.5$, respectively. We employ periodic boundary conditions, and we simulate the model for different box sizes W, from $W = 8$ to $W = 128$ (so, in accordance with $\rho = N/W^2$, the number of particles N ranging from 2^9 to 2^{15}) and for different noise strengths in $0 \leqslant \eta \leqslant 1$. We choose our observation window size l (that is, the space subset which we analyze) to be smaller than the entire system to avoid points which are near one another in space (because of periodic boundary conditions) being far apart in the Z-order curve: Since such issues arise only near the boundary, we choose $l = W/2$. Once T and coding STC or ITC are set, we indicate generically by $Q(t)$ the value of Q obtained by encoding the configuration which arises from the time interval $[t, t + T)$. We then consider the time series of Q by skipping T steps (in order to avoid information overlap) between a value of $Q(t)$ and the next one $Q(t + T)$. Finally, we time-average $\langle Q\rangle$ over $K > 10^3$ configurations, starting to collect the data after waiting for the system to reach a stationary steady state ($t > t_0 \simeq 10^4$ for the largest size). Thus

$$\langle Q\rangle = \frac{1}{K}\sum_{k=1}^{K} Q(t_0 + kT).$$

A. Setting the discretization scale *b*

We first study the effect of the cell size b at $T = 1$ (for which there is no difference between STC and ITC). As seen in Fig. 4, there is an optimal value at $b = 1$ for most of the choices of the noise intensity. This result needs some discussion. Intuitively, one could expect that the optimal value is related to the correlation length in the system. However, the perspective offered by the CID is opposite: A larger or smaller correlation length implies better or worse compression (higher or lower Q). b only sets the encoding resolution, its optimal value representing the compromise between too coarse and too refined observation. Our understanding is that, at a given noise amplitude η, the maximum of Q is found when b is of the order of the size of the smallest domains which move together. In fact, if we choose a higher value, we lose details and then information; if we choose a smaller value, we do not add information (neighboring cells hosting the same domain

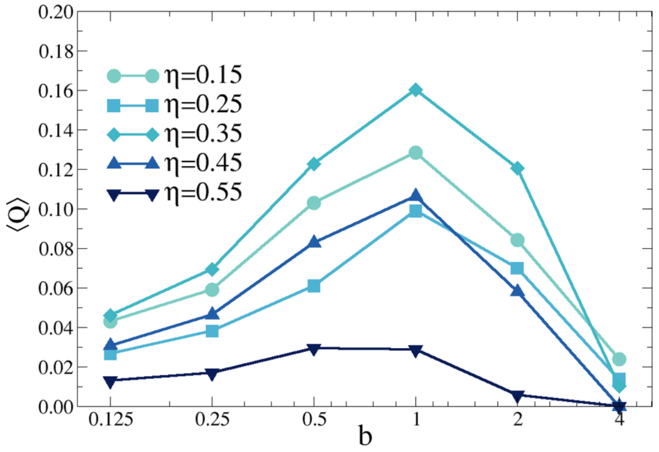

FIG. 4. The trend of Q for $T = 1$ by varying the cell size b. For finite-size systems, Q tends to trivial value 0 for $b \to 0$ (the few filled cells are randomly distributed) and $b \to l$ (each cell is filled with a 1); consequently, since $Q \geqslant 0$, there must be a maximum for some intermediate value, in this case, $b = 1$. (Data are for VM simulations with intrinsic noise; parameters are $\rho = 2$, $W = 128$, and $l = 64$.)

behave similarly and are redundant), but we add randomness to the shuffled sequences on which we normalize the CID: Both cases lead to a smaller Q. The size of these minimal domains is of the order of the interaction radius (which is $R = 1$), and such domains are present for any value of noise strength, explaining the uniformity of the optimal b seen in Fig. 4. We confirmed this hypothesis by further analysis with $R = 1$ at different speed and density values, for which we found the optimal cell size $b = 1$ again. An attempt to estimate the correlation length by looking at how compression factor scales as a function of coarse graining can be found in Ref. [11]. Here, since we want to study the effect of other parameters, in particular, the effect of the time interlacing, we fix $b = 1$ in the remainder of this paper.

B. Dependence on the time window T and on the time encoding

In Fig. 5(a) we plot the average $\langle Q \rangle$ as a function of the noise η for the ITC and STC protocols for the model with intrinsic noise and different time-window lengths T. Although the curves all have the same general shape, for $T > 1$ the values given by the ITC protocol are considerably larger than those coming from the STC one. A larger Q implies better compression, i.e., a smarter discovery of correlations in the strings. The ITC protocol also exhibits a wider dynamic range of Q (maximum value − minimum value) when parameters, such as the interleaving direction (Z-order permutation) and, most importantly, η, are changed. Finally, the ITC protocol is more sensitive to T, suggesting that it takes advantage of temporal correlations. In the following we analyze all these aspects in detail.

In Fig. 5(b), we compare step by step the "worst" result for ITC—the minimum value $Q_I^<$ over the six permutations—with the "best" result for STC—the maximum value $Q_S^>$ over the two permutations—and we find a gap $\Delta = Q_I^< - Q_S^>$ between the two encoding protocols which is always positive: This shows that the ITC protocol extracts more correlations than STC, since Δ is larger than statistical fluctuations. When T is increased, the shape of the curve Q vs η remains similar, but the values of Q obtained with the STC scheme increase slightly (mainly because of the larger statistics of substrings), while a significant increase in Q is obtained when using the ITC protocol [insets in Fig. 5(a)]. For instance, the ITC with $T = 2$ is everywhere larger than STC with $T = 64$, indicating a vastly greater sensitivity, even when $\eta \to 1$, where the particles are evenly distributed with no polarization. It is not difficult to guess why. The STC scheme preserves only space locality, so at large noise values, where there are no spatial correlations, it gives $Q \simeq 0$. In contrast, the ITC scheme preserves space and time locality and therefore exploits temporal correlations, resulting in $Q > 0$ even at high noise.

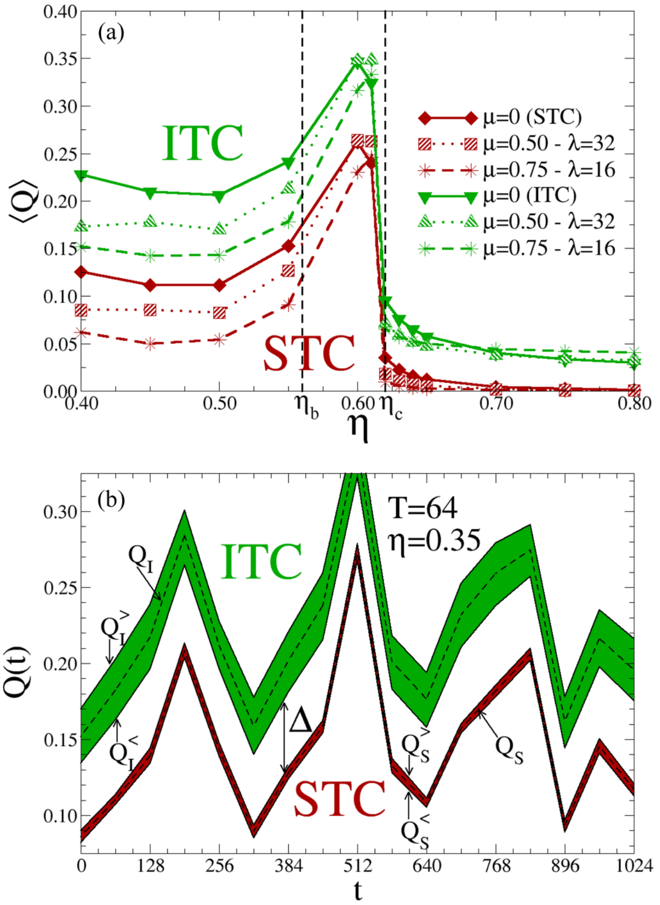

FIG. 5. (a) Q computed by the simple time-concatenation STC scheme (red curves) and by the Z order in space-time with the ITC scheme (green curves) compared with Q computed on a single configuration (black curve $T = 1$). Polarization Φ is the dashed gray line. The ITC scheme gives a Q higher than STC already at $T = 2$, and its Q increases significantly faster than STC as $T \to 64$ (insets). (b) Evolution of Q variability range due to permutations. We show that STC and ITC schemes produce, not only on average but also step by step, Q values that are always well separated. (Data are for VM simulations with intrinsic noise; parameters are $N = 32\,768$, $W = 128$, $l = 64$, and $T = 64$.)

This is also appreciated by studying how Q changes when the distance Δt between successive times in a sequence of T configurations is increased (Fig. 6). This analysis shows

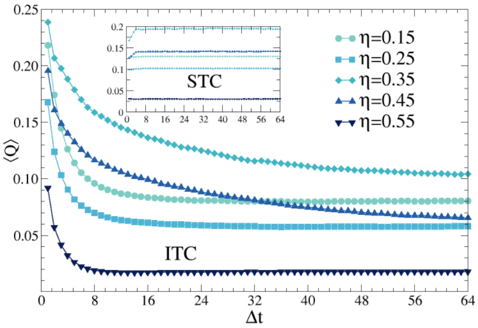

FIG. 6. Consequence for Q when the sequences of T length are built by skipping Δt time steps. The ITC scheme shows a decay of Q with varying Δt that, as we expect, slows down as the transition is approached. Differently, the STC scheme (inset) does not show meaningful information. (Data are for VM simulations with intrinsic noise; parameters are $N = 32\,768$, $W = 128$, $l = 64$, and $T = 64$.)

a striking difference between STC and ITC. STC is not particularly sensitive to Δt: It treats all configurations as independent, without respect to their proximity in time. ITC, on the other hand, displays a smooth relaxation towards an asymptotic value for large Δt. From these results we conclude that temporal interlacing (ITC) has greater sensitivity about dynamics, information that is more challenging for the STC scheme to reveal.

C. Probing the phase diagram with Q

In this section we give details about how Q describes the full complex phase diagram of the VM. We consider both variants of the VM, with intrinsic and extrinsic noise, in order to investigate the sensitivity of Q to the known differences between these two kinds of noise. Since we wish to analyze the effect of system size W and since a cubic space-time grid with power-of-2 size enables some optimization which allows us to speed up simulation, encoding, and analysis, we increase T linearly with W ($T = W/2 = 2^k$). For the larger size we simulated, we located the critical points η_b and η_c by looking by eye at the traveling band formation in running simulations, as seen in Fig. 7, where some representative simulation screenshots for selected values of η are shown. We estimate $\eta_b \simeq 0.34$ and $\eta_c \simeq 0.48$ for intrinsic noise and $\eta_b \simeq 0.56$ and $\eta_c \simeq 0.62$ for extrinsic noise. The traveling band structures that accompany the polarization transition at η_c lose coherence at smaller values of $\eta \sim \eta_b$, leading the density field to become homogeneous. At smaller values of η the density field develops strong disordered fluctuations. Figure 7 provides a detailed account of how Q computed with the ITC schema correlates with this behavior and of how it reveals more order than the STC schema.

We note that the inflection points in the Q vs η curves (or the maxima of its derivative) are able to locate the critical values η_c and η_b. Intrinsic noise is known to bear the signature of a smooth flocking transition [31], while extrinsic noise exhibits a sharper transition at a higher value of η [27].

Both behaviors are well reproduced by the Q vs η curve: In the intrinsic noise case, Q has a smooth variation close to the known value of η_c (see also the dashed curve reproducing the polarization order parameter), while in the extrinsic noise case Q has a rapid variation near η_c (which is larger). The variation of Q in the vicinity of the transitions is made clearer by looking at the derivative $|dQ/d\eta|$, shown in the insets in the left panels of Fig. 7. Remarkably, Q signals not only the polarization transition but also the other crossover present in the VM phenomenology. In particular, Q reaches a local maximum at the point marked by η^*. The fastest decrease (when reducing η) is marked by a second peak in $|dQ/d\eta|$, at a noise value η_b. This decrease is associated with the aforementioned loss of order of the density field. For even smaller values of $\eta < \eta_b$ we observe a final increase in Q with decreasing η. This is due both to the further increase in polarization and to the appearance of giant fluctuations, a well-studied phenomenon in the VM at low values of noise [29]. Such strong inhomogeneities of the density field appear as large areas with correlated values of the occupation field, contributing to an increase in Q. We observe that the steepness of the variation of $Q(\eta)$ in the proximity of η_c is stronger for larger systems, perhaps being related to the sharpness of bands, which are known to be more visible in large systems and with extrinsic noise. As can be seen in Fig. 5, these features are not a prerogative of ITC only: Albeit with a slightly weaker signal, STC and $T = 1$ analysis (from which no velocity information can be inferred) both lead to similar features. In the case of the VM, this is not surprising, since in this model, when the velocities start to align, there is a simultaneous ordering in the density field. This means that the spatial correlations of density fluctuations contain information about the state of the system and, therefore, preserving space locality only, as STC schema and $T = 1$ schema do, is enough to discriminate the phase of the system.

In the right panels of Fig. 7 we see the difference between ITC and STC as a function of η. The difference increases with N at any noise value η, with a few exceptions (we recall that at low values of noise the large correlations in the system are associated with large fluctuations). In particular, the difference between STC and ITC shows a rapid increase when crossing from above the polarization transition η_c: This fact is strictly related to the ITC's ability to retrieve information about the time correlations and to the typical slowing-down phenomena close to phase transitions. In Fig. 6 there is a clue that supports this hypothesis: The characteristic relaxation times of Q vs ΔT curves grow when η is close to critical values $\eta_c \simeq 0.48$ and $\eta_b \simeq 0.34$ suggesting precisely a power-law rather than an exponential time decay. Note that the rapid increase in the difference between ITC and STC when crossing from above the polarization transition η_c makes the ITC more sensitive in marking the polarization transition, as it implies a steeper variation of $Q(\eta)$ at that point.

D. Coping with corrupted data

The VM is a numerical model that produces trajectories for which identity, position, and velocity of each particle are exactly known at each time step. In real experiments on collective biological systems, reconstructed trajectories (both in two and three dimensions) may become corrupted in several possible ways, especially when analyzing large groups. For example, in some cases the trajectories of certain individuals

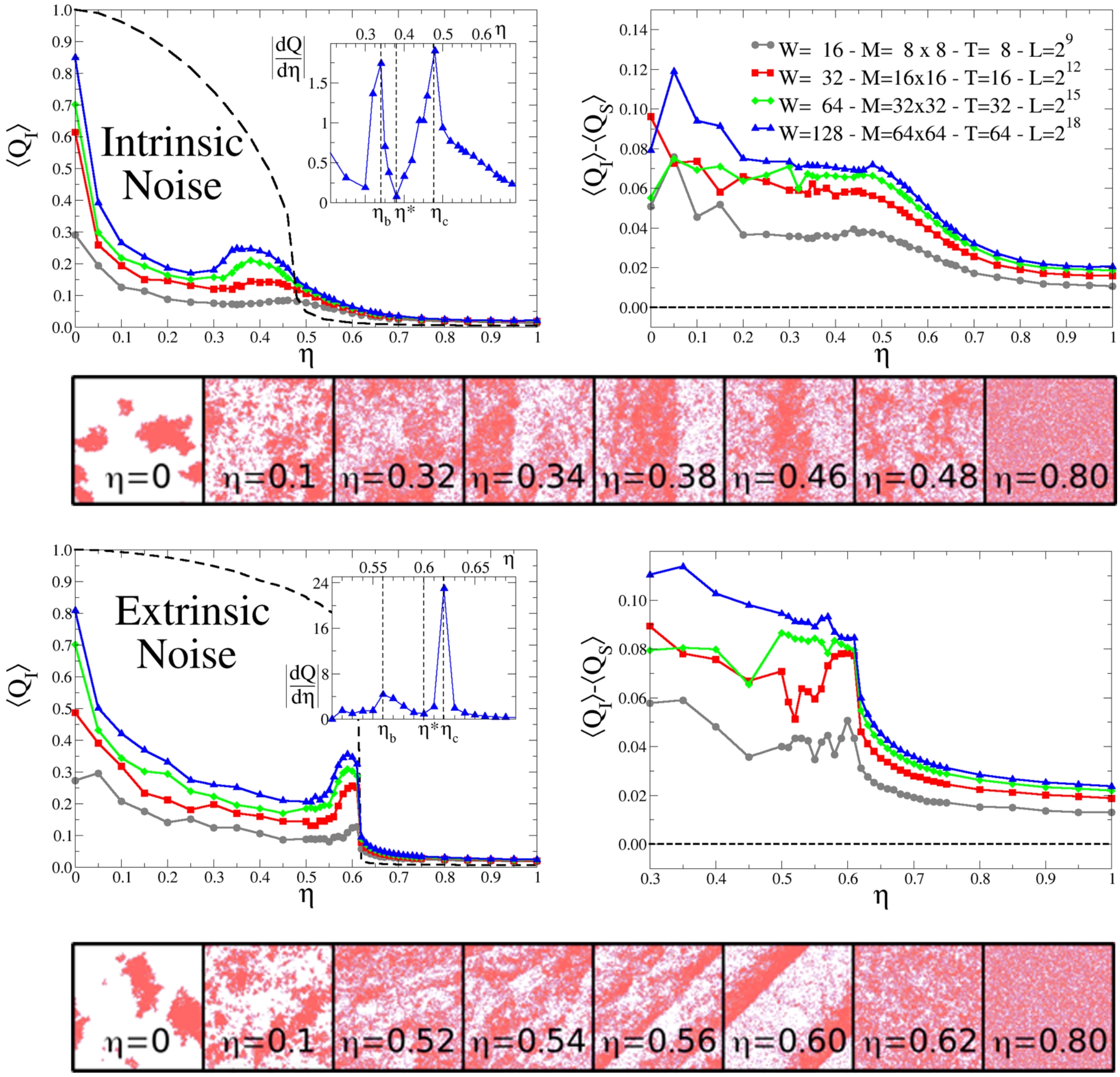

FIG. 7. Results obtained with VM simulations, both with intrinsic (top) and extrinsic (bottom) noise. Left: $\langle Q \rangle$ vs η computed with the ITC scheme at increasing system size W and time length T; dashed black curves refer to polarization Φ computed for $W = 128$. The inset shows the two peaks of $|dQ/d\eta|$ for $W = 128$ located at η_b and η_c (and the local maximum at η^*) which delimit the phase of traveling bands. The density structures are shown in the strips below. We estimate $\eta_c \simeq 0.48$ and $\eta_b \simeq 0.34$ for intrinsic noise and $\eta_c \simeq 0.62$ and $\eta_b \simeq 0.56$ for extrinsic noise. Right: Differences compared with the STC scheme with varying noise strength.

are temporarily lost, so that their positions are missing for several time steps. Thus some trajectories are interrupted, meaning that at each time step we lack information about a certain fraction of objects. This uncertainty fraction typically grows with the system density [33,34].

Here, we examine whether our analysis is sensitive enough to show data on ordering in the presence of data corruption, by simulating the interruption of the trajectories by a simple two-state Markov process. Such a process has two parameters that modulate the degree of data corruption: the fraction of missing individuals μ and the average length of a trajectory λ (see the Appendix for further details). Figure 8 shows that both STC and ITC are able to detect the transition even in the presence of strong data corruption. Since the corruption spoils correlations, this robustness of Q is nontrivial. An important point is that, with increasing corruption of data, ITC copes significantly better than STC. Not only is Q always larger for ITC than for STC, but their difference *increases* with increasing data corruption [Fig. 8(b)], apart from a few cases at small values of μ. This indicates that for real data sets, ITC will be more robust than STC; the reason seems to be that ITC better exploits time correlations to recover the information which is lost because of “vanished” particles.

IV. CONCLUSIONS

We have studied a measure based on data compression and applied it to the Vicsek model, a nonequilibrium active system which describes collective behavior in biological systems. Motivated by the desire to establish a framework for the analysis of actual data on collections of living things, we adopted a crude encoding method, based on a binary coarse graining of the positional information. We do not directly feed the velocity information to the encoding, despite the crucial role of

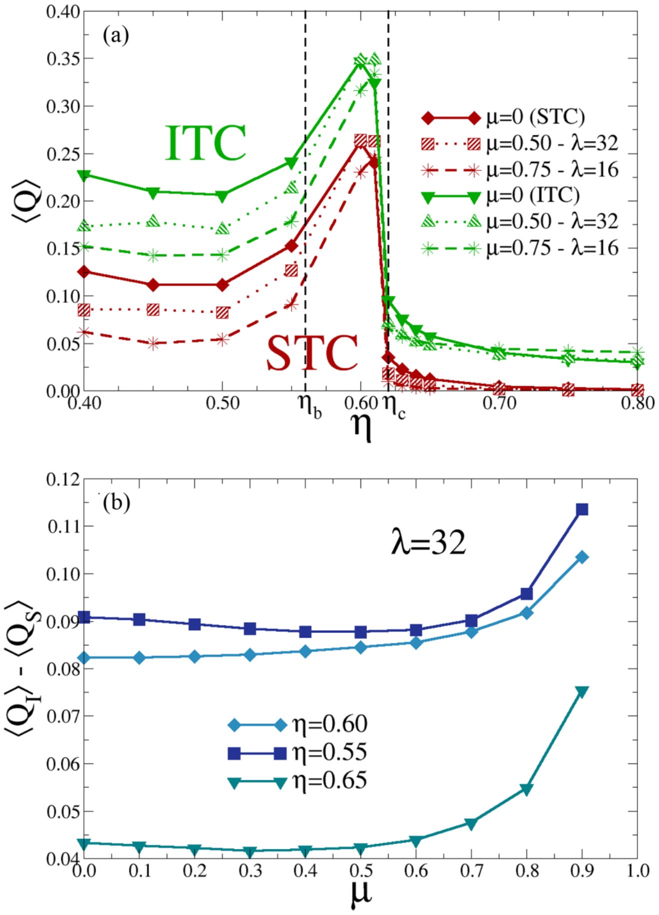

FIG. 8. (a) The effect of data corruption on $\langle Q\rangle$ vs η. (b) Dependence on μ of the difference between $\langle Q_I\rangle$ and $\langle Q_S\rangle$ for some value of noise strength η. (Data are for VM simulations with extrinsic noise; parameters are $N = 32\,768$, $W = 128$, $l = 64$, and $T = 64$.)

velocity alignment in the VM. Numerical results show that the CID is able to capture the order-disorder transition normally described by the velocity order parameter, demonstrating that it can exploit the density-velocity coupling present in the VM and in many nonequilibrium systems. This result is promising for future analysis of real data, as the positional information is easily obtained in experiments.

One could, in principle, explore other, more refined, ways of encoding the physical information, by changing, and expanding, the alphabet of the compressed string according to other local properties of the system (including velocities). However, we have seen that a larger alphabet makes the space of possible strings larger: Therefore—at fixed string length L—it reduces the ability of the compressing scheme to exploit correlations. Of course, a larger alphabet also makes strings longer (at constant configuration size), which implies longer time for CID computation: In our implementation, such a time increases slightly more slowly than linearly with L (and therefore with the number of bits needed to encode additional features). Given the results obtained with our simple encoding method, we deem that more complex encodings are not necessary in general.

Extending our approach to systems in three (or more) dimensions is straightforward, since the Z-ordering curve has an obvious generalization in any dimension (and it is usually faster and simpler to implement than the Hilbert curve). The method is not constrained to regular cubic lattices with $T = m = 2^n$ (we have chosen to do so just to simplify computational work). In fact, regardless of the size of the observation window and of grid features, once the cells have been defined, it is sufficient to sort them according to their Z value. Effects of the aspect ratio of the simulation box are taken into account by averaging on the permutations of the coordinates, as illustrated in this paper.

Our results show that preserving locality in both space and time in the encoding (ITC), rather than in space only (STC), is important. ITC always extracts more information than STC, and most importantly, the difference in the performance is robust (or even increases) in the presence of corruption of the data sets. This result is particularly important if one wants to apply the method to actual biological data. A free parameter of this procedure is the size of the coarse-graining cells b: Here, we have calibrated it looking for the value giving maximum compression (larger CID), which was associated with the interaction radius and therefore did not vary appreciably when changing the other parameters. We believe that it should behave similarly in other applications.

We conclude by describing some future directions of our work. The use of compression-based tools is particularly promising for the study of the response to perturbations in collective biological systems. The fluctuation-dissipation theorem (FDT), connecting the unperturbed correlations of a given observable to the linear response of the system to a given small external perturbation, is particularly simple in equilibrium, where the observables involved are dictated by the Hamiltonian. In the case of flocks, swarms, and other biological systems, one has to exploit one of the many recipes for nonequilibrium generalization of the FDT [4]. In all such recipes, one needs to know which of the relevant variables are conjugated to the perturbation; however, in the case of biological experiments, it is not at all clear which variables of the systems are perturbed in the presence of a given external stimulus. Recent advances in nonequilibrium statistical physics provide a possible way out: For nonequilibrium steady states the response to perturbations can always be expressed in terms of correlations involving observables conjugate with respect to a specific observable, the “stochastic entropy” [35]. We conjecture that the observable Q investigated here is closely related to it. Our analysis, in particular, shows that Q is coupled to many relevant degrees of freedom in the system and therefore is a promising candidate for a general approach to response in biological systems.

ACKNOWLEDGMENTS

We thank Irene Giardina and Stefania Melillo for fruitful discussions. This publication has been produced as a part of the Italian-Israeli scientific cooperation project “IDA-CReBS”: A.C., A.P., and M.V. were supported by Ministero degli Affari Esteri e della Cooperazione Internazionale, and D.L. was supported by the Israel Ministry of Science and Technology (Grant No. 3-15150). A.C. and M.V. were supported by European Research Council (Advanced Grant “RG.BIO” No. 785932). A.P. acknowledges the financial support of Regione Lazio through the "Progetti Gruppi di Ricerca" (Grant No. 85-2017-15257) and from the MIUR PRIN 2017 (Project No. 201798CZLJ). D.L. was supported by the Israel Science Foundation (Grant No. 1866/16) and the U.S.-Israel Binational Science Foundation (Grant No. 2014713). P.M.C., D.L., and S.M. were supported by the National Science Foundation Physics of Living Systems program (Grant No. 1504867). P.M.C. was partially supported by the Materials Research Science and Engineering Center program of the National Science Foundation (Award No. DMR-1420073).

APPENDIX

1. The LZ77 ALGORITHM

We here illustrate how the LZ77 algorithm works and derive the corresponding definition for the CID, with an example: We try to encode a sequence of characters in a list of "longest previous factors" (LPFs). We represent a LPF by a pair of integers (p, l) which are the "instructions" to retrieve the original sequence: *"Print l symbols starting from the pth character already written; if $l = 0$, just print p directly."* To decode, we must go through the list of LPFs in order and, for each of them, follow the instructions.

Consider the binary sequence of length $L = 16$

$$0101100111001110,$$

and read it from left to right. At the beginning, no characters have been printed yet so, surely, the first two LPFs are (“0″, 0) and (‘‘1″, 0) (both have $l = 0$; in this case, p is the character to be printed, and we emphasize it using quotation marks). Next we have a substring "01" that we have already met: The LPF is then (1,2) (copy two characters starting from position 1). The next substring already met is "10," and it is encoded by LPF (2,2), followed by (3,3) to encode "011." Finally, we see that the remaining substring "1001110" is obtained by copying seven characters starting from position 5, so the last LSF is (5,7). It does not matter whether the sequence currently available is shorter than seven characters and incomplete, as the full subsequence becomes available during printing. The number of bits $\mathcal{L}$ needed to encode this list of $C = 6$ LPFs can be estimated by the following argument. Let (p_i, l_i) be the ith LPF and $b_i \simeq \log_2 p_i + \log_2 l_i$ be the number of bits needed to encode it: In this way the length of the original sequence L and the compressed binary length $\mathcal{L}$ are given by

$$L \simeq \sum_{i=1}^{C} l_i, \quad \mathcal{L} = \sum_{i=1}^{C} b_i.$$

Since $p_i < L$, we have $b_i \leqslant (\log_2 L + \log_2 l_i)$, and $\mathcal{L}$ must be bounded by

$$\mathcal{L} \leqslant C \log_2 L + \sum_{i=1}^{C} \log_2 l_i.$$

Now, we use Jensen's inequality for concave functions

$$\sum_{i=1}^{C} \log_2 l_i \leqslant C \log_2 \left(\frac{1}{C} \sum_{i=1}^{C} l_i \right) = C \log_2 \frac{L}{C}.$$

After simple algebraic manipulation, we obtain an expression for the CID as in Ref. [10]

$$\frac{\mathcal{L}}{L} \leqslant \frac{C \log_2 C + 2C \log_2 (L/C)}{L}.$$

The CID of our example ($C = 6$, $L = 16$) is 2.03. For effective compression we must consider longer sequences; for example, consider an $L = 128$ string obtained by replicating the sequence considered above ($L = 16$) eight times. In this case we must add one more LPF, (1,112), so, since $C = 7$ and $L = 128$, we find CID $\simeq 0.61$, and we have almost halved the number of bits needed to represent the sequence.

2. Data corruption

In order to mimic real data corruption or degradation, we proceed as follows. We associate with each particle i a Boolean random variable $b_i \in \{0, 1\}$ which evolves under the action of a two-state Markov chain with transition probability $P(b \to b') = P_{bb'}$:

$$p = P_{10} = 1 - P_{11}, \quad q = P_{01} = 1 - P_{00}.$$

At each simulation step we apply the rules of this Markov process to evolve b_i stochastically. In this way, we are able to modulate the degree of data corruption by tuning two parameters: (i) the fraction of missing individuals μ and (ii) the average length of a trajectory, λ. In particular, we consider or ignore particle i of the data set according to the value of b_i: When $b_i = 0$, the particle i is removed from the data set until b_i returns to 1. It is easy to prove that the invariant measure ρ_b ($\rho_0 = \mu$) and the typical length (the average life span) λ of a trajectory depend on p and q in the following way:

$$\mu = \rho_0 = 1 - \rho_1 = \frac{p}{p+q}, \quad \lambda = \frac{\sum_{l=1}^{\infty} l P_{11}^{l}}{\sum_{l=0}^{\infty} P_{11}^{l}} = \frac{1-p}{p}.$$

Then, by setting p and q to appropriate values

$$p = \frac{1}{1+\lambda}, \quad q = \frac{1-\mu}{\mu} p$$

and by starting with a configuration of $\{b_i\}_{i=1,N}$ already in equilibrium according to the invariant measure $\sum_i b_i / N \simeq \rho_1$, we can simulate data corruption by varying μ and λ.

[1] W. H. Zurek, *Complexity, Entropy and the Physics of Information* (CRC, Boca Raton, 2018).

[2] G. J. Chaitin, *Information, Randomness & Incompleteness: Papers on Algorithmic Information theory*, World Scientific Series in Computer Science Vol. 8 (World Scientific, Singapore, 1990).

[3] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).

[4] U. M. B. Marconi, A. Puglisi, L. Rondoni, and A. Vulpiani, Phys. Rep. **461**, 111 (2008).

[5] C. E. Shannon, Bell Syst. Tech. J. **27**, 623 (1948).

[6] E. Vogel, G. Saravia, F. Bachmann, B. Fierro, and J. Fischer, Phys. A (Amsterdam) **388**, 4075 (2009).

[7] O. Melchert and A.K. Hartmann, Phys. Rev. E **91**, 023306 (2015).

[8] J. L. Lebowitz and H. Spohn, J. Stat. Phys. **95**, 333 (1999).

[9] J. M. Parrondo, C. Van den Broeck, and R. Kawai, New J. Phys. **11**, 073008 (2009).

[10] S. Martiniani, P. M. Chaikin, and D. Levine, Phys. Rev. X **9**, 011031 (2019).

[11] S. Martiniani, Y. Lemberg, P. M. Chaikin, and D. Levine, Phys. Rev. Lett. **125**, 170601 (2020).

[12] D. Benedetto, E. Caglioti, and V. Loreto, Phys. Rev. Lett. **88**, 048702 (2002).

[13] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, Phys. D (Amsterdam) **180**, 92 (2003).

[14] M. Henkel, H. Hinrichsen, and S. Lübeck, *Non-Equilibrium Phase Transitions - Volume 1: Absorbing Phase Transitions* (Springer, New York, 2008).

[15] M. E. Cates and J. Tailleur, Annu. Rev. Condens. Matter Phys. **6**, 219 (2015).

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 2012).

[17] D. Sheinwald, A. Lempel, and J. Ziv, IEEE Trans. Commun. **38**, 341 (1990).

[18] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Phys. Rev. Lett. **75**, 1226 (1995).

[19] J. K. Parrish and W. M. Hamner, *Animal Groups in Three Dimensions* (Cambridge University Press, Cambridge, 1997).

[20] G. Grégoire, H. Chaté, and Y. Tu, Phys. D (Amsterdam) **181**, 157 (2003).

[21] J. Toner, Y. Tu, and S. Ramaswamy, Ann. Phys. (Amsterdam) **318**, 170 (2005).

[22] J. Toner and Y. Tu, Phys. Rev. Lett. **75**, 4326 (1995).

[23] Y. Tu, J. Toner, and M. Ulm, Phys. Rev. Lett. **80**, 4819 (1998).

[24] J. Toner and Y. Tu, Phys. Rev. E **58**, 4828 (1998).

[25] M. Aldana, V. Dossetti, C. Huepe, V. M. Kenkre, and H. Larralde, Phys. Rev. Lett. **98**, 095702 (2007).

[26] G. Baglietto and E. V. Albano, Phys. Rev. E **80**, 050103(R) (2009).

[27] G. Grégoire and H. Chaté, Phys. Rev. Lett. **92**, 025702 (2004).

[28] A. P. Solon, H. Chaté, and J. Tailleur, Phys. Rev. Lett. **114**, 068101 (2015).

[29] F. Ginelli, Eur. Phys. J.: Spec. Top. **225**, 2099 (2016).

[30] H. Chaté, F. Ginelli, G. Grégoire, and F. Raynaud, Phys. Rev. E **77**, 046113 (2008).

[31] M. Nagy, I. Daruka, and T. Vicsek, Phys. A (Amsterdam) **373**, 445 (2007).

[32] G. M. Morton, *A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing* (IBM, Ottawa, 1966).

[33] A. Attanasi, A. Cavagna, L. Del Castello, I. Giardina, A. Jelić, S. Melillo, L. Parisi, F. Pellacini, E. Shen, E. Silvestri, and M. Viale, IEEE Trans. Pattern Anal. Mach. Intell. **37**, 2451 (2015).

[34] A. Cavagna, S. Melillo, L. Parisi, and F. Ricci-Tersenghi, IEEE Trans. Pattern Anal. Mach. Intell. **43**, 1394 (2021).

[35] T. Speck and U. Seifert, Europhys. Lett. **74**, 391 (2006).