Short-range interactions versus long-range correlations in bird flocks

Andrea Cavagna, Lorenzo Del Castello, Supravat Dey, Irene Giardina, Stefania Melillo, Leonardo Parisi, and Massimiliano Viale

Phys. Rev. E **92**, 012705 — Published 6 July 2015

DOI: 10.1103/PhysRevE.92.012705

Short range interactions vs long range correlations in bird flocks

Andrea Cavagna[a,b,c], Lorenzo Del Castello[a,b], Supravat Dey[a,b], Irene Giardina[a,b,c], Stefania Melillo[a,b], Leonardo Parisi[a,b,d], and Massimiliano Viale[a,b]

[a] *Istituto Sistemi Complessi, Consiglio Nazionale delle Ricerche, UOS Sapienza, 00185 Rome, Italy*
[b] *Dipartimento di Fisica, Università Sapienza, 00185 Rome, Italy*
[c] *Initiative for the Theoretical Sciences, The Graduate Center,*
365 Fifth Avenue, New York, NY 10016 USA and
[d] *Dipartimento di Informatica, Università Sapienza, 00198 Rome, Italy*

Bird flocks are a paradigmatic example of collective motion. One of the prominent traits of flocking is the presence of long range velocity correlations between individuals, which allow them to influence each other over the large scales, keeping a high level of group coordination. A crucial question is to understand what is the mutual interaction between birds generating such nontrivial correlations. Here we use the Maximum Entropy (ME) approach to infer from experimental data of natural flocks the effective interactions between individuals. Compared to previous studies, we make a significant step forward as we retrieve the full functional dependence of the interaction on distance, and find that it decays exponentially over a range of a few individuals. The fact that ME gives a short-range interaction even though its experimental input is the long-range correlation function, shows that the method is able to discriminate the relevant information encoded in such correlations and single out a minimal number of effective parameters. Finally, we show how the method can be used to capture the degree of anisotropy of mutual interactions.

I. INTRODUCTION

Large groups of animals - such as bird flocks, fish schools and insect swarms - display a remarkable degree of collective coordination. Several experimental studies in the last decade quantified the spontaneous emergence of global order [1–3], the presence of strong behavioral correlations between individuals [4, 5], and the swift transfer of information through the group [2, 6–8]. Such findings stimulated a multi-disciplinary interest in this kind of systems. On the one hand, animal groups can be considered as instances of active matter [9, 10], and can be expected to display some of the non trivial properties observed and predicted for several living, soft and granular active systems on the micro-scale. On the other hand, there are important features that make animal groups more complicated to understand. First, the way individuals coordinate with one another are not only determined by physical mechanisms, as for rods or hard discs, but also (and often mainly) by exquisitely biological processes (including cognitive). As a consequence, any speculation about the nature of mutual interactions in a group cannot be taken for granted. Besides, animal aggregations form large, but not infinitely large groups: they are not in the thermodynamic limit, but rather live in an intermediate regime where finite size effects can be important [5]. Understanding collective animal behavior therefore implies understanding what is the nature of interactions, what are the effective features of such interactions that are relevant on the scale of natural groups, and how they determine the collective properties that we observe.

One of the most intriguing features of collective animal motion is the presence of long-range correlations. The correlation function of the velocity fluctuations has been found to be long-range both in polarized groups such as bird flocks [4] and in disordered ones, such as insect swarms [5]. These results suggest that, rather than ordering, what is truly characteristic of collective behavior in biological systems is the ability of individuals to correlate changes in behavior and influence each other over the large scales. It is therefore important to understand what are the features of the interactions granting such strong correlations.

We know from statistical physics that short-range interactions are sufficient to produce spontaneous symmetry breaking and system-level coordination. Models of self-propelled particles [3, 11–15] and hydrodynamic flocking theories [10, 16] have shown numerically and analytically that also in active systems short-range interactions can produce global ordering and long-range correlations. Indeed, there now seems to be some consensus in the field of collective animal behavior that interactions are short-ranged [18]. However, long-range interactions do exist in nature, so we cannot rule them out *a priori*. Moreover, to create long-range correlations out of short-range interactions one normally needs some special conditions: either there is a continuous spontaneously broken symmetry (Goldstone theorem), or the system is in the scaling region of a critical point. From a biological perspective, one could legitimately object that a reasonable long-range interaction is a better explanation of long-range correlations than some arcane physical theorem, not to mention criticality. For example, birds' vision is likely to span the entire size of a flock. It is worth noticing that, despite the short-range consensus, it has been recently proposed that a *long-range* interaction is at the basis of flocking behavior [19]. Hence, the notion that short-range interactions rule collective behavior, albeit reasonable, is still far from being an established fact, even in those systems that have been most intensively studied experimentally.

The recent access to large scale empirical data on animal groups has considerably advanced our understanding of the problem. However, a direct and unbiased proof of whether interaction is short- or long-ranged has been lacking so far. Significant results have been obtained by fitting biological models to the data [20–23]. Yet, the problem with model fitting is that it may be tricky to distinguish the intrinsic properties of the system under investigation from the a priori ingredients of the model used to fit the data. Alternatively, the interaction has been assessed by using some structural proxy of it. For example, in [25] the authors measured experimentally the anisotropy in the spatial distribution of the neighbors around a given bird, and found it to decay over a range of a few individuals (~ 7). Since this anisotropy can only be determined by the interaction, the authors concluded that interactions must be short range and decaying over approximately the same range. However, through this kind of structural proxies one does not attain *direct* access to the interaction.

In this paper we follow a different approach and use the maximum entropy method [24] to infer the interactions directly from the data. The philosophy of this method is different from standard model fitting in that, as we will discuss, the model it designs for the system is dictated by the available experimental observables and it is not assumed a priori. For bird flocks, we started this program in [26], with encouraging results. Using a very simple experimental input, we inferred the effective number of individuals each bird is interacting with, and the average strength of such interaction. Hence, in [26] it was *assumed* a step-like shape of the interaction, in order to keep the mathematical complexity to a minimum. Here, we make a significant step forward and derive the *full functional dependence on distance* of the effective alignment interaction between individuals. We call this function $J(n)$, where n is the topological distance between birds, i.e. their order of neighborhood [25]. We find that $J(n)$ decays exponentially, on scales much smaller than the system size, indicating that alignment interaction within a flock is short-range. The experimental input of our calculations is the velocity correlation function, which is long-range. We show however that much of the information captured by the correlation function is redundant and only correlations on a short scale are sufficient to retrieve the interaction $J(n)$. Hence, not only can we infer the effective interaction, but we only need a small number of local experimental measurements. Thanks to the new method we are also able to study the angular dependence of the interaction with respect to the direction of motion of the flock and shed some light on the anisotropic spatial arrangement of neighboring birds found in past experimental studies [25, 28].

The paper is organized in the following way. In Sec.II we introduce and describe the Maximum Entropy approach, we outline the mathematical structure of the computation and apply it to the case of flocks. In Sec.III we show the results of the computation for the flocking events in our dataset; we compare them with previous work [26]; and we further generalize the method to capture the possible angular anisotropy of the interactions. In Sec.IV we discuss what are the effects of changing the number of experimental input parameters in the calculation and show that a fair result is achieved when the inferred interaction does not depend on the number of input parameters anymore. Finally, in Sec.V we discuss our results.

II. MAXIMUM ENTROPY APPROACH TO FLOCKS

Collective phenomena and ordering transitions have been widely studied in condensed matter. From the perspective of statistical physics, one usually knows the microscopic interactions between particles, and wants to predict their large scale properties. When dealing with biological systems we often face the opposite situation. We have access to collective observables through experiments, but have scarce knowledge on the effective interactions generating them. The problem is in this case an *inverse* one: to build a microscopic statistical model starting from the macroscopic data. As mentioned in the Introduction, several approaches have been developed to deal with this task, from model fitting to Bayesian inference [29]. Here we consider the Maximum Entropy (ME) approach. This method was originally established by E. T. Jaynes in 1957 [24] and has strong connections with classical statistical physics. In the last decade, it has been widely used to describe the collective behavior of biological networks, from neural assemblies, to amino acids in proteins, biochemical and genetic networks and flocks of birds [26, 27, 30–32, 34–41].

The main idea of the ME method is to build the least structured statistical model - the maximum entropy model - which is consistent with a given set of measured observables. In this section we explain how to construct a ME model and how the method can be applied to the case of birds flocks. Before doing this, however, we would like to make two remarks on the method, which are useful to understand its philosophy, appreciate its results and evaluate its performance.

i) As compared to other approaches, the ME principle has the remarkable feature that it does not rely on a priori assumptions on the system under study; this means that ME does not assume any form of the microscopic interactions (at variance with model fitting). This does not mean that the ME does not make approximations in the description of the system; in fact it does, but we have a way to control and evaluate them. As we shall see in detail in this Section, the kind of model we get from the ME approach crucially depends on the experimental observables we consider as input. Were we able to perform good measurements of many observables we would retrieve in an accurate way the full probability distribution of the micro-states of our system. This is not however what happens in real experiments, where typically only a few quantities can be measured, and not always in a robust statistical way. What we know is that, given some observables, the ME model is the one that describes them best with the least number of assumptions. In this sense, the effective interactions appearing in a ME model only come from the experimental behavior of the system. Besides, once we construct a ME model based on some experimental input, we have a way to test its predictive power. We can, for example, use the model to predict quantities other than the ones used to build it, and compare to experiments. More systematically, we can compute the predictive gain acquired when providing new experimental input and assess the information content of the ME model. What happens in some cases is that a few experimental input observables give significant gains, and adding further experimental input makes weak progress. We will discuss an example of this procedure in Sec.IV.

ii) If we consider as input observables quantities that are - at least on certain timescales - stationary, the ME method provides a static ME model. As we shall see, in this case the ME distribution has the form of a Boltzmann measure, which is particularly useful from a mathematical point of view to perform computations. This does not mean, however, that the system is in equilibrium, nor that the ME distribution is an equilibrium one. In fact the system can have an arbitrary off-equilibrium dynamics. In this case the ME method simply captures the effect of this dynamics on the statistical distribution of a given set of observables. What is relevant to us in this paper is that this distribution encodes how certain degrees of freedom effectively interact due to the microscopic behavior of the system. We note that the ME approach is not bound to produce static Boltzmann-like measures. If we consider as input observables time dependent quantities (such as multi-point time correlation functions), the resulting ME model will consist in a time dependent distribution [40, 42–44]. Computations and inference of effective interactions can be in this case much more complicated. For polarized self-propelled systems we showed that, as long as the network of positions does not rearrange too fast (which is the case of natural flocks [51]), static and dynamic ME models give a very similar inference of the interaction parameters [40].

A. The general ME scheme

Consider a system whose micro-state at any instant of time is described by a set of variables $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_N\} \equiv \mathbf{X}$. When the size of the system N becomes large, the space of the $\mathbf{X}$ increases exponentially and it is experimentally impossible to directly sample and reconstruct the probability distribution $P(\mathbf{X})$. On the contrary, it is usually possible to accurately measure aggregate observables, which require less statistics. Let us assume that we can measure several observables $f_1(\mathbf{X}), f_2(\mathbf{X}), ...f_M(\mathbf{X})$, and let us denote their experimental averages by $\langle f_1\rangle_{\rm expt}, \langle f_2\rangle_{\rm expt}, ..., \langle f_M\rangle_{\rm expt}$, respectively. The maximum entropy (ME) method consists in finding the most random probability distribution $P(\mathbf{X})$ that is consistent with the observed experimental data. The distribution must therefore satisfy the following constraint:

$$\langle f_\mu\rangle_{\rm expt} = \langle f_\mu\rangle_{\rm P} \ , \tag{1}$$

for all $\mu = 1, 2, ..., M$, and where $\langle f_\mu\rangle_{\rm P} = \sum_{\mathbf{X}} P(\mathbf{X}) f_\mu(\mathbf{X})$ denotes the expectation value computed using the probability distribution $P(\mathbf{X})$. Many distributions satisfy Eq. (1). The maximum entropy principle [24] aims to find the one which has as little structure as possible, i.e. is the most random, so that one can derive the minimal consequences of the experimental observations on $\langle f_\mu\rangle_{\rm expt}$. As a measure of randomness of a given distribution we consider its Shannon entropy [45, 52],

$$S[P] = -\sum_{\mathbf{X}} P(\mathbf{X}) \log P(\mathbf{X}). \tag{2}$$

In order to get the desired probability distribution, we then need to maximize $S[P]$ under the constraints given by Eq. (1). Besides the experimental constraints, there is an additional constraint, namely that the probability distribution should be normalized $\sum_{\mathbf{X}} P(\mathbf{X}) = 1$. This is equivalent to say that we add to our list of observables an extra function, the constant $f_0(\mathbf{X}) = 1$. This constraint maximization problem can be solved with the Lagrange multiplier method [55] by finding the optimum of the generalized entropy function,

$$S[P; \{\lambda_\nu\}] = S[P] - \sum_{\mu=0}^{M} \lambda_\mu \left(\langle f_\mu\rangle_{\rm P} - \langle f_\mu\rangle_{\rm expt}\right), \tag{3}$$

where each Lagrange multiplier λ_μ is associated with a constraint equation. Maximizing $S[P; \{\lambda_\nu\}]$ with respect to $P(\mathbf{X})$, we get

$$P(\mathbf{X}) = \frac{1}{Z(\{\lambda_\nu\})} \exp\left[-\sum_{\mu=1}^{M} \lambda_\mu f_\mu(\mathbf{X})\right], \tag{4}$$

where $Z(\{\lambda_\nu\})$ enforces the normalization and is obtained optimizing with respect to λ_0

$$Z(\{\lambda_\nu\}) = \exp(1 + \lambda_0) = \sum_{\mathbf{X}} \exp\left[-\sum_{\mu=1}^{M} \lambda_\mu f_\mu(\mathbf{X})\right]. \tag{5}$$

Using Eq.(4) the generalised entropy (3) can be written as a function of the Lagrange parameters only,

$$S(\{\lambda_\nu\}) = \log Z(\{\lambda_\nu\}) + \sum_{\mu=1}^{M} \lambda_\mu \langle f_\mu\rangle_{\rm expt}. \tag{6}$$

One can now easily optimize with respect to λ_μ to recover the original constraint equation (1),

$$-\frac{\partial \log Z(\{\lambda_\nu\})}{\partial \lambda_\mu} = \sum_{\mathbf{X}} P(\mathbf{X}) f_\mu(\mathbf{X}) = \langle f_\mu\rangle_{\rm expt}. \tag{7}$$

As we can see from Eq.(4), the maximum entropy distribution has the form of a Boltzmann distribution $P(\mathbf{X}) = \exp(-\beta H(\mathbf{X}))/Z$ with an effective "Hamiltonian" $H(\mathbf{X}) = \sum_{\mu=1}^{M} \lambda_\mu f_\mu(\mathbf{X})$ and temperature $k_B T = 1$. As we previously discussed, this does not mean that the system we are looking at is in equilibrium nor that this Hamiltonian is the true microscopic Hamiltonian of the system (if it exists). Nevertheless, one must not forget that the optimal values of the Lagrange parameters, through Eq.(7), enforce consistency with experimental data. They describe the effect of the microscopic dynamics on the statistics of the input observables. In this sense, they represent effective interactions and mirror the structure of the microscopic behavior of the system through the filter of our experiments. In this respect, the choice of which experimental observables to consider as input of the method is very important: the more representative the set of $\{f_\mu(\mathbf{X})\}$ is of the collective behavior of the system, the more predictive the ME model (4) turns out to be, the more informative the effective parameters are on the microscopic features of the system that determine its behavior at the collective scale.

Keeping these considerations in mind, to investigate the nature of interactions in our system, we need to select a good set of experimental input observables, compute the corresponding ME model, investigate its predictive content, and finally look at the structure of the effective ME interactions.

From a mathematical point of view, to compute the ME model we need to solve Eqs. (4) and (7). This means computing $Z(\{\lambda_\nu\})$, which represents the partition function of the Boltzmann-like distribution (4), a problem we are fairly well-equipped to deal with (at the level of schemes and approximations) in statistical physics. There is, however, a further obvious difficulty: Z is not a number, but a function of the Lagrange parameters, and must be computed for any possible value of the $\{\lambda_\nu\}$. This is the essence of the inverse problem. In most cases this is a hard step, which is achieved numerically. For flocks, which are very polarized groups, one can resort to a high-order expansion and compute $Z(\{\lambda_\nu\})$ analytically. Once this function is known, we can fix the values of the Lagrange parameters by enforcing the constraint, i.e. by optimizing Eq.(7).

Interestingly, we note that the generalized entropy (7) is related to the likelihood of the experimental data $\mathcal{L}(\{\lambda_\nu\})$, i.e.

$$\log \mathcal{L}(\{\lambda_\nu\}) = \langle \log P \rangle_{\rm expt} = -\log Z(\{\lambda_\nu\}) - \sum_{\mu=1}^{M} \lambda_\mu \langle f_\mu \rangle_{\rm expt} = -S(\{\lambda_\nu\}) \tag{8}$$

Hence, optimizing the generalized entropy (which - as can be shown - corresponds to a minimum in the parameters' space) is equivalent to maximize the log-likelihood of the experimental data.

B. ME distribution for flocks

Let us now apply the ME method to flocks of birds. The first step is to identify the set of variables that defines the microstate of the system under investigation (i.e. the variables $\{\mathbf{X}\}$ of the previous section). We are interested in the interaction that is responsible for the alignment of the directions of motion of the birds. Hence we consider as microscopic variables the orientation vectors, $\vec{s}_i \equiv \vec{v}_i/|\vec{v}_i|$, where $\vec{v}_i$ is the velocity of bird $i = 1, \ldots, N$.

The second step is to select a set of observables, function of the variables $\vec{s}_i$, whose experimental value will be used to constrain the probability distribution of the orientation vectors, $P(\{\vec{s}_i\})$. The standard way to proceed is to use moments (i.e. correlations) of this distribution, $\langle \vec{s} \rangle, \langle \vec{s} \cdot \vec{s} \rangle, \ldots$. As one can easily see from Eq.(4), each one of these m-points correlations will generate m-points interaction terms in the effective Hamiltonian. One could naively think that the more correlations we consider, the better the corresponding ME model. In fact this is not true, for several reasons. On the experimental side, the larger is m the larger is the statistics needed to get good experimental estimates (in terms of number of measurements and sample size). Thus, using large m-point correlations typically enhances the experimental noise. From a more conceptual point of view, not all correlations are in general equally important. By considering too many of them we can introduce redundant information and risk to overfit the parameters. The most economic prescription is therefore to use up to the minimum m-point correlation that allows to predict the $m+1$-point correlation.

Previous studies have shown that in flocks (as in other collective systems [30]) the use of pairwise interactions ($m = 2$) allows to accurately predict the 4-points correlations [26, 27]. We then focus on pairwise correlations. In a flock of birds we can in principle define the mutual correlation of the flight directions $C_{ij} = \vec{s}_i \cdot \vec{s}_j$ for any single pair of individuals. However, these quantities wildly fluctuate in time and never reach a steady state. The reason is obvious: birds are not on a fixed lattice; they move in space so to change position with respect to each other. Therefore, the mutual distance of i vs j changes in time; but any reasonable social force (i.e. interaction) will depend on the *relative* distance between individuals, rather than on their *absolute* identity. Hence, distance, rather than identity, should be used as a label. The experimental proof of that is that the correlations computed as a function of distance are stable in time over appreciable intervals. Besides, they exhibit a non-trivial scale free dependence, signature of the collective behavior of the flock [4], which makes them a very good choice for our purposes.

We therefore consider as our experimental input observable the two-point correlation function,

$$\hat{C}(n; \{\vec{s}_i\}) = \frac{1}{N} \sum_{i,j=1}^{N} \vec{s}_i \cdot \vec{s}_j \; \delta(k_{ij} - n) \; . \tag{9}$$

This quantity is the average correlation between a bird and its n^{th} nearest neighbour (we use the hat notation to distinguish this full correlation from the connected one - see below). Compared to previous studies [4], we measure the correlation as a function of the topological distance (i.e. order of neighborhood), n, rather than of the metric distance, r [25]. In Eq.(9) k_{ij} is the topological distance of bird j relative to bird i: if j is the first nearest neighbour of i, $k_{ij} = 1$; if j is the second nearest neighbour of i, $k_{ij} = 2$; and so on (k_{ij} is nonsymmetric). This implies $\sum_{i,j} \delta(k_{ij} - n) = N$, and this is why the normalization in (9) is easier than in its metric counterpart [4, 46].

As explained in the previous section, the ME method consists in finding the probability distribution $P(\{\vec{s}_i\})$ that maximizes the entropy $S[P]$ under the constraint that the distribution reproduces the experimental observables (9), which in our case read

$$\langle \hat{C}(n; \{\vec{s}_i\}) \rangle_{\text{expt}} = \langle \hat{C}(n; \{\vec{s}_i\}) \rangle_{\text{P}} \; . \tag{10}$$

This constrained maximization is achieved by introducing one Lagrange multiplier, $J(n)$, for each experimental quantity that we are fixing, $\hat{C}(n)$. It is convenient to define $J(n)$ as an intensive parameter so that, in the notation of sec.II A, $J(n)$ corresponds to λ_μ/N, and $\hat{C}(n)$ corresponds to f_μ. As we have explained, the distribution obtained in this way has the form of an exponential of the product of the Lagrange multipliers times the observables

$$P(\{\vec{s}_i\}) = \frac{1}{Z} \, e^{N \sum_n J(n) \hat{C}(n)} \frac{1}{Z} \, e^{\sum_{ij} J(k_{ij}) \, \vec{s}_i \cdot \vec{s}_j} \; , \tag{11}$$

where Z is the normalizing partition function and $J(k_{ij}) = \sum_n J(n) \delta(n - k_{ij})$. The probability distribution (11) corresponds to the effective Hamiltonian,

$$H = -\sum_{ij} J(k_{ij}) \, \vec{s}_i \cdot \vec{s}_j \; . \tag{12}$$

Therefore, the (discrete) function $J(n)$ represents the strength of the effective alignment interaction between pairs of birds at topological distance n. Once we solve the ME model and we compute $J(n)$, we can therefore investigate the nature of such interaction, how it decays in distance and understand whether it is short- or long-ranged.

Inferring the full function $J(n)$ is a significant step forward compared to our previous ME calculations [26, 40], where we assumed a step-like shape of the interaction. By doing that we only had to infer two parameters, intensity and range of the interaction, so that we had no information about the *form* of the interaction. For this reason, the question of short vs long range interaction in [26] was addressed in a rather indirect way, namely by checking that the interaction range did not scale with the system size. Here, on the contrary, we will be able to calculate directly how the interaction decays and to see explicitly that it is short range.

C. Maximization of log-likelihood

To retrieve the interaction function $J(n)$ we need to solve the ME equations enforcing the constraints (10) or, equivalently, maximize the log-likelihood of the data. In our case the log-likelihood function Eq.(8) is given by

$$\log \mathcal{L} = \langle \log P(\{\vec{s}_i\}) \rangle_{\text{expt}} = -\log Z[J(n)] + N \sum_n J(n) \langle \hat{C}(n; \vec{s}_i) \rangle_{\text{expt}}. \tag{13}$$

We therefore need to compute the partition function $Z[J(n)]$ and then perform the maximization with respect to $J(n)$. This is a non-trivial program. There are however a few tricks we can exploit to facilitate the task. We outline here the main steps, details can be found in the Appendixes.

- The expression of the effective Hamiltonian can be simplified further. We can indeed rewrite H by introducing the symmetrized interaction matrix J_{ij},

$$H = -\sum_{ij} J_{ij} \, \vec{s}_i \cdot \vec{s}_j \tag{14}$$

where $J_{ij} \equiv [J(k_{ij}) + J(k_{ji})]/2$. Interestingly, Eq. (14) describes the Hamiltonian of an Heisenberg model on a network, whose topology is described by the interaction matrix J_{ij}. In the strongly ordered phase - as flocks are - this model can be solved using a well known low temperature expansion, the spin-wave approximation (see Appendix). As a result, Z can be computed analytically and is entirely given in terms of the eigenvalues $\{a_k\}$ of discrete Laplacian matrix $A_{ij} = \delta_{ij} \sum_k J_{ik} - (1 - \delta_{ij})J_{ij}$, giving

$$Z[J(n)] = -\sum_{k>1} \log a_k + \sum_n J(n) \tag{15}$$

- In principle we need to consider all possible values of mutual distances $n = 1 \cdots N$, and optimize over N distinct Lagrange parameters. This number can be however severely reduced. It turns out (see next section) that correlations $C(n)$ for $n > n_{max}$ are redundant and do not improve the computation. Thus, all the sums appearing in Eq. (13) can be extended only up to n_{max}. Besides, one can 'bin' the integer values of topological distances n in discrete intervals of size Δn, much as one would do with real values of the metric distance (see Appendix B). In this way the number of effective variational parameters can be reduced even further, speeding up the maximization procedure. The expression of the log-likelihood then becomes

  $$\log \mathcal{L} = \sum_{k>1} \log a_k - N\Delta n \sum_n{}' J(n)(1 - \langle \hat{C}(n) \rangle_{\text{expt}}). \tag{16}$$

  where the primed sum indicates that we are summing over discrete bins, up to n_{max}.

- The derivatives of the second term of the log-likelihood with respect to the $J(n)$ are trivial. However, differentiating the partition function is far less trivial, as the eigenvalues a_k are very complicated functions of the $\{J(n)\}$. Luckily we can use perturbation theory (see Appendix C) and derive the exact expressions for the derivatives w.r.t. to $J(n)$.

In this way we finally get the ME equations

$$\langle \hat{C}(n) \rangle_{\text{expt}} = 1 - \frac{1}{N\Delta n} \sum_{k>1} \frac{1}{a_k} \frac{\partial a_k}{\partial J(n)} = 1 - \frac{\text{Tr}[A^{-1}\gamma(n)]}{N\Delta n} \ , \tag{17}$$

where the matrix γ is given by

$$\gamma_{ij}(n) = \frac{1}{2}\delta_{ij} \left[\sum_m (\delta(k_{im} - n) + \delta(k_{mi} - n)) \right] - \frac{1}{2}(1 - \delta_{ij})(\delta(k_{ij} - n) + \delta(k_{ji} - n)). \tag{18}$$

These equations can be exploited to efficiently maximize the log-Likelihood (16) numerically (see Appendix D for details on the numerical procedure) and find, for each value of n, the optimal $J(n)$. The results of this procedure are discussed in the next Section.

III. RESULTS

A. Short range interactions vs long range correlations

Let us summarize the procedure explained so far. We considered a set of experimentally measured observables, the velocity correlation functions Eq.(9), and built the ME distribution consistent with these observables. This distribution is expressed in terms of effective alignment interactions between individuals, whose dependence in mutual distances is described by the function $J(n)$. The ME allows us to retrieve $J(n)$ by maximizing the log-likelihood, given the experimental input $\langle \hat{C}(n) \rangle_{\text{expt}}$.

Let us now discuss the results of this procedure. We used an experimental dataset of 22 flocking events (see Table I and Appendix E). Data were obtained from stereoscopic experiments in the field: large flocks of starlings (from hundreds to thousands birds) were filmed with high resolution stereo-cameras and - thanks to innovative computer vision techniques - individual 3D tracking was performed [4, 25, 49]. Given the difficulty of the problem, this dataset represents to date the largest experimental dataset on large animal groups moving in three dimensions.

For each event, we measured the correlations $\langle \hat{C}(n) \rangle_{\text{expt}}$ and used them as input for the ME computation. The resulting $J(n)$ is plotted in Fig.1, for two distinct flocks. As we can see from the figure, the interaction function (red line) clearly decays to zero on a topological scale of few (order ten) individuals. To fully appreciate the result and its consequences we also plotted in the same figure the connected correlation function (blue line), which measures the decay of correlations between birds. So far we always considered the *nonconnected* velocity correlation function $\hat{C}(n)$, Eq.(9). Flocks are however in the ordered phase (they have nonzero polarization), hence $\hat{C}(n)$ does not decay to zero. This is simply a consequence of the emergence of long-range order: all birds fly on average in the same direction and there is a trivial contribution of the center of mass motion to the full correlation function. For this reason, if we want to describe how correlations decay, we need to consider the *connected* correlation function, which is defined by using the velocity fluctuations

$$C(n) = \frac{1}{N} \sum_{i,j} \delta\vec{s}_i \cdot \delta\vec{s}_j \, \delta(k_{ij} - n) \ , \tag{19}$$

where $\delta\vec{s}_i = \vec{s}_i - (1/N)\sum_k \vec{s}_k$. $C(n)$ is basically the full correlation minus the order parameter squared and measures how much individual deviations from the average motion are correlated with each other.

For very large systems (i.e. in the thermodynamic limit) $C(n)$, unlike $\hat{C}(n)$, decays to zero for large distances, signaling the physical fact that fluctuations must be uncorrelated when their distance tends to infinity. In finite systems, however, the behaviour of the connected correlation function depends on the nature of the correlation in the system. In systems with short-ranged correlations, namely systems where the correlation length ξ is always smaller than the systems size, $C(n)$ decays to zero as in an infinitely large system; this means that not only the function reaches the zero axis at a distance of the order of the correlation length, but it also stays zero beyond this distance and the correlation function does not depend much on the system's size. Bird flocks, however, have been found to belong to a very different class, namely systems with long-range correlations, also called scale-free systems [4]: in this case, the correlation length ξ scales with the system's size and the infinite size form of the correlation function is a power law. This fact that has several implications; first, in a scale-free system the zero of the connected correlation function, which is the best proxy of a correlation length, scales with the systems size; second, because of the definition of connected correlation, where fluctuations are calculated by subtracting the spatial average, the function $C(n)$ crosses the zero axis in correspondence to the correlation length, without leveling to zero after this point (for an extended discussion of this point see [4]). Only for very large systems one would see the power-law for of the correlation function.

In Fig. 1 a,c we can compare the behavior of $C(n)$ (the input) with the inferred effective interaction strength, $J(n)$ (the output). What we find is that, in contrast with the correlation, the interaction $J(n)$ is very much *short-ranged*. The difference between $C(n)$ and $J(n)$ is quite striking (Fig. 1, left). We find that $J(n)$ decays exponentially with the topological distance (Fig. 1 b,d),

$$J(n) = J_0 \, e^{-n/n_c} \ , \tag{20}$$

where the decay constant n_c provides a measure of the interaction range. The mean value of n_c over all 22 analyzed flocks is,

$$n_c = 8.0 \pm 0.5 \quad \text{(std error)} \ , \tag{21}$$

to be compared with the estimate $n_c = 6.5 \pm 0.9$ (std error) given in [25] using spatial structure as a proxy of the interaction. Plots of $J(n)$ for several other analyzed events are displayed in Fig.2, while the values of n_c for all events can be found in Table I. In all cases the interaction decays exponentially and the interaction range n_c is much smaller than - and not dependent on - the system's size N (see Table I). In terms of metric distances, in all cases these ranges correspond to distances much smaller than the extension of the flock (and well below its shorter dimension) - see Table I. We therefore find that the effective alignment interaction in starling flocks is short-ranged [50].

We notice another interesting aspect of Fig.2: these plots all have the same scale on the abscissa, meaning that in all flocks the interaction decays over a similar range of the topological distance n. But these flocks have significantly different densities, therefore if we wanted to plot J as a function of the physical, metric distance r we would need widely different scales. This is yet another demonstration of the previously discovered fact [25, 26] that interaction in bird flocks is based on topological, rather than metric distance [50].

B. Comparison with the step-interaction case

As mentioned in the Introduction, a first, simpler, maximum entropy computation on bird flocks was performed by some of us in [26]. Let us now compare the present approach with the one of [26], and discuss the present step forward compared to previous results.

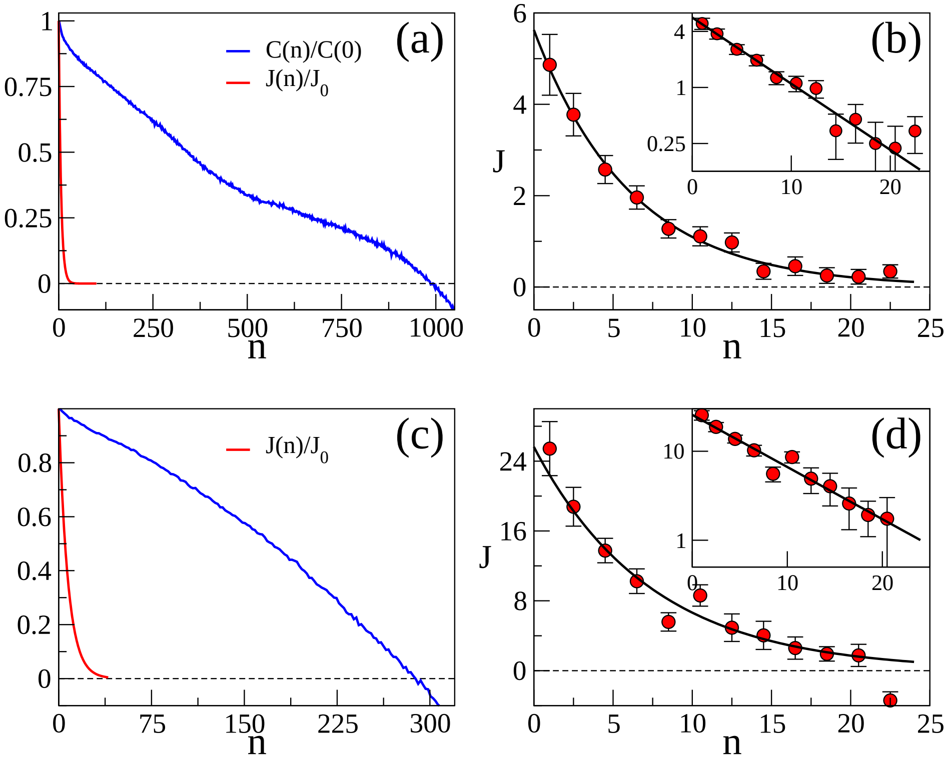

FIG. 1. **Left** Connected correlation function $C(n)$ compared to the interaction $J(n)$ (both normalized by their $n = 0$ value to be displayed on the same scale). **Right** Close-up of the interaction, $J(n)$. The full line is an exponential fit to the data (see text). Inset: Semi-log plot of the same quantity. **Top.** Event 31-01, $N = 2126$; $J_0 = 5.63$, and $n_c = 6.11$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$.

The experimental input used in [26] was also related to velocity correlations. However, it was not the correlation as a function of distance $\hat{C}(n)$, as in the present paper. Rather, we considered a single scalar, $\hat{C}_{\rm int}$, describing the *average* degree of correlation between a bird and its interacting neighbors

$$\hat{C}_{\rm int} = \frac{1}{N}\sum_i \frac{1}{n_c}\sum_{j\in i}^{n_c} \vec{s}_i \cdot \vec{s}_j, \tag{22}$$

where the sum is carried out over the first n_c neighbours of i. We can recast this quantity in the language of the present paper by noting that,

$$\hat{C}_{\rm int} = \frac{\sum_{i,j=1}^{N} \vec{s}_i \cdot \vec{s}_j \,\Theta(k_{ij} - n_c)}{\sum_{i,j=1}^{N} \Theta(k_{ij} - n_c)}, \tag{23}$$

where $\Theta(x)$ is the Heaviside step function. By using the formalism that we developed above, it is easy to see that this construction is equivalent to assume that the interaction function has a step-like behavior, being constant up to neighbour n_c and zero beyond that, namely,

$$J(n) = J_0\ \Theta(n_c - n). \tag{24}$$

In [26] J_0 - the (average) strength of the interaction - naturally appeared as the Lagrange multiplier associated to $C_{\rm int}$, so that the entropy was maximized w.r.t. it. On the other hand, n_c - the width of the step-like interaction - was *not* the Lagrange multiplier of any given observable, so it remained in the likelihood even after maximization w.r.t. J_0 and was determined through a maximum likelihood principle. This means that the calculation of [26] in fact assumed some parameter-dependent form of the model (namely the step interaction form (Eq. (24))), and did not exclusively rely on entropy maximization.

How does the calculation of [26] compare to the one we developed above? First of all, we see from Fig. 2 that the old step-interaction is always compatible with the new exponential interaction, so there is a nice consistency between

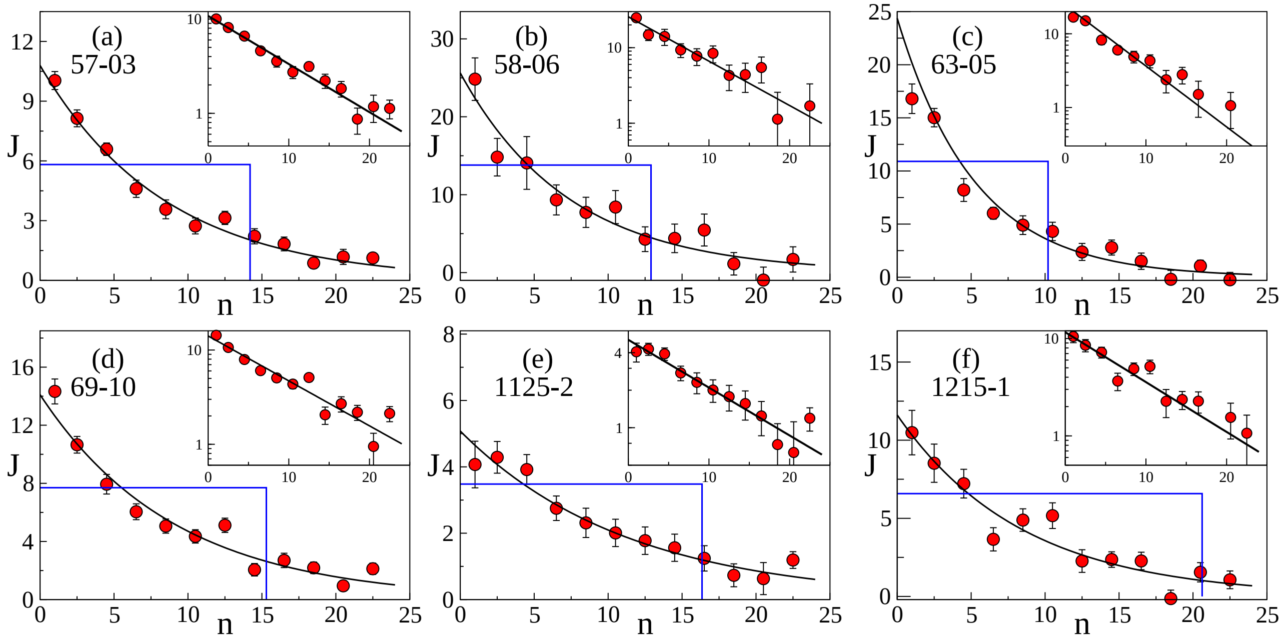

FIG. 2. The alignment interaction strength $J(n)$ for six events different from those in Fig.1: event 57-03 (a), event 58-06 (b), event 63-05 (c), event 69-10 (d), event 20111125-2 (e), and event 20111215-1 (f) (See the Table. I for the details of the events). Circular symbols (red) are the result of the ME method, while black solid lines are the exponential fit to the data. Step functions (blue) are the interaction strengths computed using the ME method described in [26]. Insets are semi-log plots of $J(n)$ for the respective events. All of them show reasonably clear exponential decays.

the two cases. For a more quantitative comparison, let us call $J_0^{\rm step}$ and $n_c^{\rm step}$ the strength and the range of the interaction of the step model of [26], and $J_0^{\rm exp}$ and $n_c^{\rm exp}$ the parameters of the exponential fit of the $J(n)$ that we calculated in the present work. It is reasonable to expect two things:

1. the total interaction strength, that is $\sum_n J(n)$, should be the same in the two cases. From this condition we get $n_c^{\rm exp} J_0^{\rm exp} = n_c^{\rm step} J_0^{\rm step}$;

2. if we interpret $w(n) = J(n)/\sum_m J(m)$ as the (normalized) weight of the n^{th} neighbour, the average of n, i.e. $\sum_n w(n)n$ should be the same in the two cases.

These two conditions give,

$$n_c^{\rm exp} = n_c^{\rm step}/2 \ , \tag{25}$$

$$J_0^{\rm exp} = 2 J_0^{\rm step} \ . \tag{26}$$

The data in Fig. 3 indicate that these two relations are indeed satisfied. Notice that the fact that the step-like parameter $n_c^{\rm step}$ is twice as large as the exponential decay range can (at least partially) explain the discrepancy between $n_c^{\rm step}$ (which was found to be $= 21.6$ in [26]) and the previous estimate of the interaction range given in [25] ($n_c = 6.5 \pm 0.9$).

C. Longitudinal vs transverse interaction

In natural flocks the distribution of neighbors around a given individual was found to be anisotropic [25]. This suggests that there might be a certain degree of anisotropy in the interactions between birds. We can use the ME approach to investigate this question. We know that the more detailed the experimental input we use, the more detailed the corresponding ME model will be. In the previous section we discussed how increasing the amount of experimental information can lead to an increase in knowledge about the interactions: using only $\hat{C}_{int}$ [26] allows inferring an effective interaction range and strength, while using the correlation function $\hat{C}(n)$ allows inferring the full dependence of interaction on distance. In the same way, to probe the angular dependence of interactions we now

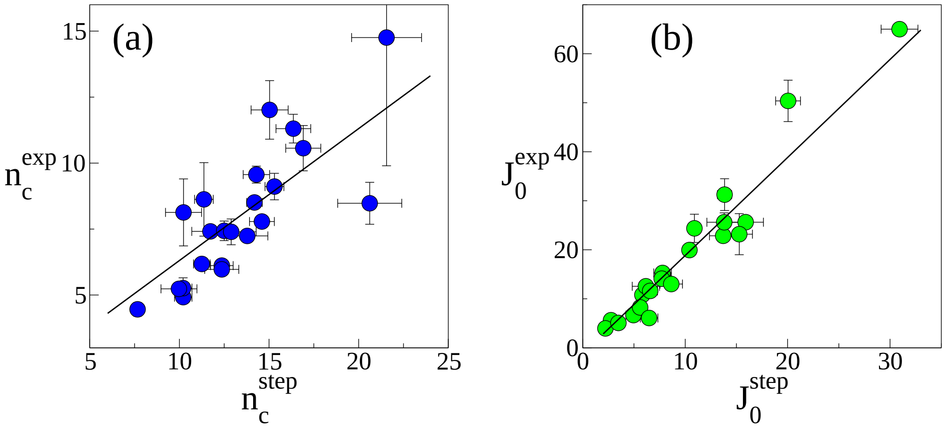

FIG. 3. (a) Interaction range n_c and (b) interaction strength J_0: Comparison of the step model vs the present work; the full lines are the predictions of Eq. (25) and Eq. (26).

consider correlation functions, which not only depend on distance, but also the on angle with respect to the direction of motion.

Given a bird, i, we partition the space around it into two sectors, the *longitudinal* one and the *transverse* one: consider a neighbor j of i, and let θ_{ij} be the angle formed by $\vec{r}_{ij}$ (the vector joining i to j) and the flock's direction of motion $\vec{V}$; then j is in the longitudinal sector of i if $|\cos(\theta_{ij})| > 1/2$; otherwise it falls into the transverse sector (this relationship is symmetric). Notice that with this definition the two sectors have the same $3d$ volume. We then define the longitudinal (L) and transverse (T) correlation functions, which are simply the average correlations in their relative sectors,

$$\hat{C}^{L,T}(n) = \frac{\sum_{i,j} \vec{s}_i \cdot \vec{s}_j \delta(k_{ij} - n)\Theta(\pm|\cos(\theta_{ij})| \mp 1/2)}{\sum_{i,j} \delta(k_{ij} - n)\Theta(\pm|\cos(\theta_{ij})| \mp 1/2)} \tag{27}$$

where $\Theta(x)$ is the Heaviside step function. When computing these correlations on flocks data, we find that the transverse correlation is slightly but systematically larger than its longitudinal counterpart at small topological distances: the percentage of times where $C^T(n = 1) > C^L(n = 1)$ is 64% (over all frames and events), and is above 50% in 91% of events. Starting from these new observables, one can apply the maximum entropy method as explained in the previous sections and get a ME distribution with effective Hamiltonian;

$$H = -N \sum_n \left[p^L(n) J^L(n) \hat{C}^L(n) + p^T(n) J^T(n) \hat{C}^T(n) \right] , \tag{28}$$

Here, the Lagrange multipliers $J^L(n)$ and $J^T(n)$ represent the alignment interaction strengths of a bird with its n-th neighbour in the longitudinal and transverse direction respectively. The $p^{L,T}$ are the fraction of neighbors that lie in longitudinal and transversal sector and are defined as $p^{L,T}(n) = (1/N) \sum_{i,j} \delta(k_{ij} - n)\Theta(\pm|\cos(\theta_{ij})| \mp 1/2)$. These quantities of course satisfy the relation $p^L(n) + p^T(n) = 1$. We note that, despite the constrains:

$$\hat{C}(n) = p^L(n)\hat{C}^L(n) + p^T(n)\hat{C}^T(n), \tag{29}$$

the link between $J(n)$ and $J^{L,T}(n)$ is not trivial. In this case we match at the same time $\hat{C}^L(n)$ and $\hat{C}^T(n)$, in the isotropic case $\hat{C}(n)$ only. However there are many different combinations of $\hat{C}^L(n)$ and $\hat{C}^T(n)$ that correspond to the same global $\hat{C}(n)$ (Eq. (29)). This means that we can have many combinations of $J^{L,T}(n)$ consistent with the same $J(n)$. Special cases occur only when the correlation functions of the two sectors are the same $\hat{C}^L(n) = \hat{C}^T(n)$ (in this case $J^L(n) = J^T(n) = J(n)$) or when there are no neighbors in one of the sectors (if $p^T(n) = 1$ then $J^T(n) = J(n)$ and $J^L(n)$ is indeterminate, and viceversa – as it should be).

To find the $J^{L,T}(n)$ consistent with experimental data we proceed along the lines explained in the previous sections. Also in this case the Hamiltonian (28) can be recast in an Heisenberg-like form, which allows computing analytically the partition function to get an explicit expression of the log-likelihood in terms of $J^L(n)$ and $J^T(n)$ (see Appendix B). The transverse and longitudinal interaction functions can then be retrieved by maximizing the log-likelihood.

The result is shown in Fig. 4, where we plot the values of $J^T(n)$ and $J^L(n)$ for $n = 1$ and $n = 2$ for all the analyzed flocking events. The interaction between nearest neighbors in the transverse direction is detectably stronger than

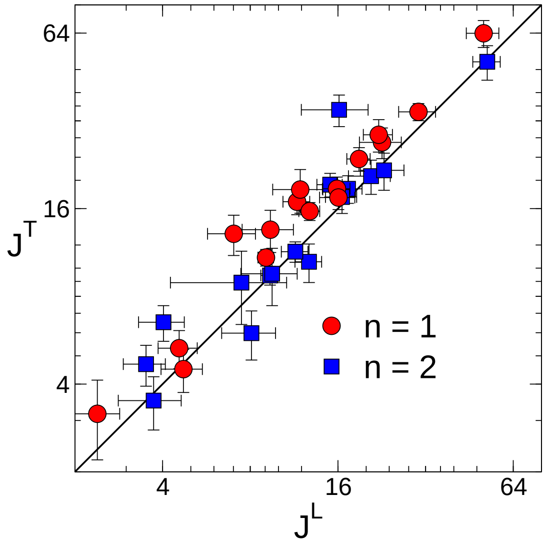

FIG. 4. Log-log plot of J^L vs J^T for $n = 1$, and $n = 2$ for all the analyzed flocks. The full line is the identity. The computation of both correlations and inferred interactions in the anisotropic case requires a larger statistics, because only half of birds pairs are on average used to get $J^{L,T}(n)$ and $\hat{C}^{L,T}(n)$. For this reason, events where the size is too small or that are too short in time are too noisy and have been not included in the analysis (events 77-07, 72-02, 1214-4-1,1214-4-2, 58-07 - see Table I)

that in the longitudinal direction. On an average, $J^T(n = 1)$ is 20% larger than $J^L(n = 1)$. More precisely, $J^T(1)$ and $J^L(1)$ are linearly correlated with Pearson correlation coefficient $\rho > 0.99$, the best linear fit giving $J^T(1) = (1.208 \pm 0.058)J^L(1)$ (statistical confidence p-value $< 1.0e-6$). This anisotropic character of the interaction is very short-ranged, though, as it already disappears by the second nearest neighbor, $J^T(n = 2) \simeq J^L(n = 2)$ (Fig. 4).

The anisotropy that we find is not strong, but it is interesting. Let us discuss more in details its possible origins and consequences.

First, remember that we are studying the *alignment* interaction, hence our result tells us that a bird is more keen to align its direction of motion with the neighbor on the side, rather than with that directly in the front. One may speculate that this is due to the fact that misalignment with a side neighbor has more severe consequences (in terms of collision) than that with someone along the direction of motion. On the other hand, for what concerns *speed* control one would expect the opposite: a stronger interaction in the longitudinal direction would be more useful to avoid bumping into each other. Some recent progress has been made in working out the speed interaction in flocks [27]; it would therefore be interesting to extend the present calculation to the case of speed.

Even though the anisotropy concerns the directional degrees of freedom (the velocities) it can have an impact on the spatial arrangement of neighbors. One can argue that individuals who better coordinate flight directions tend to keep the same mutual distance, and consistently maintain their neighborhood relationship. In this respect, our result is consistent with the finding of [25], where it was found that the closest neighbors of a bird are more easily found in the transverse than in the longitudinal direction (i.e. the nearest neighbors are typically on the side rather than in the direction of motion). Understanding how interactions between flight directions are related to the spatial structure of the group is a complex problem. There could be positional attraction-repulsion forces between birds, which we did not consider in our ME analysis, and that can influence the spatial arrangement of individuals (see [15, 53] for a discussion in numerical models). Recent analysis [41] however suggest that in systems with topological interactions velocity alignment has an important role in the structure, which is why our result can help to understand this issue.

Finally, a word of caution is required. There is an important anisotropy present in polar active systems, which is a consequence of symmetry breaking and dynamics and is not due to anisotropic microscopic interactions. Flocks are polarized groups, as such velocity fluctuations orthogonal to the global velocity are much stronger than longitudinal fluctuations, due to the presence of soft modes (see Appendix A). In self-propelled systems this causes anisotropic diffusion exponents and a non trivial scaling of correlations in the large scale hydrodynamic regime [17]. Natural flocks exhibit their collective behavior on much shorter scales - both in terms of sizes and time [7]. Still, it might be that the anisotropy captured by the ME model in part describes the effect of the microscopic anisotropic diffusion on the scale of the experimental observations.

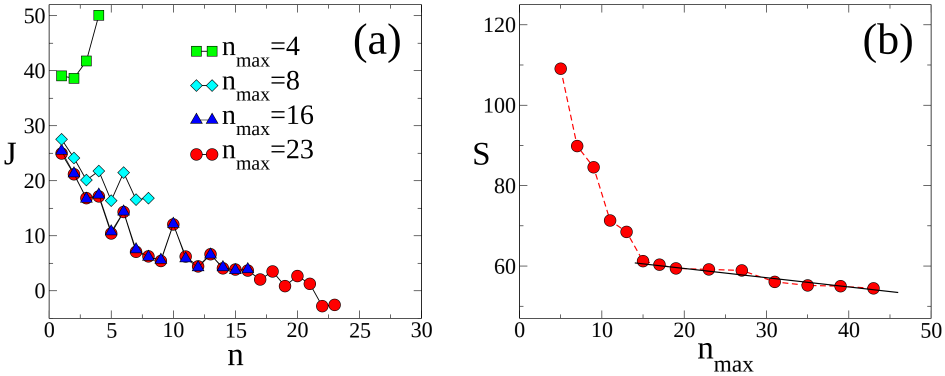

FIG. 5. (a) The interaction strength $J(n)$ is plotted for different n_{max} (event 21-06). With increasing $n_{\max}$, the form of $J(n)$ saturates. (b) Entropy S vs $n_{\max}$ for the same event. In the large n regime the entropy decays very weakly; in this regime we are merely fitting the noise.

IV. DEPENDENCE ON THE NUMBER OF INPUT VARIABLES

When we introduced the ME approach in Sec.II we briefly discussed the role of the number of input experimental observables. The more observables we consider, the more detailed the corresponding ME model. Indeed, the number of parameters that we infer through the method is equal to the number of experimental observables that we use to constrain the entropy maximization. In principle, increasing the amount of experimental input should lead to a more complete knowledge of the effective interactions in the system (see e.g. Sec.III B, III C). This is not however always true. There are cases where the relevant information is captured by a small number of observables, and one just needs these few observables to get a very effective description of the system. Our case is precisely of this kind, and offers a very nice example where the predictive power of a ME model can be clearly quantified.

In our work we use the correlation function $\hat{C}(n)$ as reference observable. The topological distance n between two birds can go up to $n_{\max} = N$. This means that the correlation function (9) is a set of N numbers, so that we should in principle use N Lagrange multipliers, $J(n)$ with $n = 1, \dots, N$, to maximize the entropy. The very long-range nature of $C(n)$ seems to suggest that there is indeed information to be exploited in this whole function, up to the maximum possible values of the topological distance. In fact, the situation is very different.

What we find is that if we use values of $\hat{C}(n)$ up to a maximum distance $n_{\max}$, the inferred interaction stabilizes for $n_{\max} \ll N$. To see this we maximized the entropy for different numbers of Lagrange multipliers, that is we calculated $J(n)$, with $n = 1, \dots, n_{\max}$, for different values of $n_{\max}$ (Fig. 5a). What we see is that for very small $n_{\max}$ the function $J(n)$ is unstable, so that the whole interaction changes drastically when increasing $n_{\max}$. However, when $n_{\max}$ becomes large enough the full interaction $J(n)$ stops depending on $n_{\max}$ and the only effect of feeding more correlations and adding new parameters to the calculation is to obtain negligible and noisy couplings. This means that beyond a certain distance, the ME calculation simply refuses to switch on any more couplings, even though the long-range correlation function that we feed as an input still seems ripe of information at that distance. This is an indication that the ME method works with remarkable economy.

The role of $n_{\max}$ can be understood also at the level of the entropy. The value of the entropy as a function of $n_{\max}$ after maximization tells us how much information we gain [45] by adding more experimental data ($\hat{C}(n)$) and by inferring more parameters ($J(n)$). We can see from Fig. 5b that the entropy decays very fast up to a certain $n_{max} \simeq 15$, and then the decay becomes slower and linear. This means that we are gaining real information for $n_{\max} \leq 15$, but after that we are simply fitting the noise and there is no more useful information to be gained by increasing $n_{\max}$. For infinitely large N, that is for infinitely accurate experimental averages, we expect the large n weak decrease of the entropy to become a real plateau, signifying that there is really nothing to gain (not even in terms of noise fitting) by adding more parameters than those really required by the short-range interaction.

To better understand the role of the entropy and fully appreciate the meaning of the change of behavior displayed in Fig. 5 we need a Bayesian analysis. If we want to infer a good model, a model that tells us something about the behavior of the system, our purpose is not simply to fit well the data. Rather, our goal is to find the *minimal* set of parameters able to reproduce the experimental data. Within a Bayesian framework this goal can be mathematically formalized. Let us call $P(n_{\max}|D)$ the probability of a model with $n_{\max}$ parameters, given a certain dataset D. It can be shown that [29]

$$P(n_{\max}|D) = P(D|n_{\max})V(n_{\max}) \sim e^{-S(n_{\max})} e^{-\alpha n_{\max}} \ . \tag{30}$$

The first term in the r.h.s. is the maximized likelihood, i.e. the probability of getting the data with a model with n_{max} parameters, and is given by the exponential of the ME entropy (Eq. (13)). The second term, which is called Occam factor, $V(n_{\max})$, is equal to the ratio between the posterior accessible volume in the space of parameters and the prior accessible volume [29]. Typically, the Occam factor decays exponentially with the number of parameters, $V(n_{\max}) \propto e^{-\alpha n_{\max}}$.

Hence, in general when we increase the number of parameter $n_{\max}$ of the model we have a trade-off: on one hand it improves the fit, hence it increases the likelihood; on the other hand, it decreases the Occam factor. Because of this trade-off, when the number of parameters increases beyond a certain value, the suppressing contribution of the Occam factor compensates the decay of the entropy, and therefore the growth of the likelihood. For this reason $P(n_{\max}|D)$ reaches a maximum for a finite value of parameters, $n_{\max} = n_{\max}^{opt}$ Fig. 6(b).

Unfortunately, the Occam factor depends on the *prior* probability of the parameters, which is always an obscure thing. For this reason the position of this maximum is not clearly defined. However, it is possible to show that this fact produces only a small ambiguity in the location of $n_{\max}^{opt}$. First of all, the contribution of the Occam factor to $P(n_{\max}|D)$ depends only logarithmically on the prior probability: major changes in the prior leads to a small change in Occam factor. Moreover, when we increase $n_{\max}$ after we reach the optimal value, the slope of entropy changes suddenly: in fact, we move from the regime $n_{\max} < n_{\max}^{opt}$, where adding each new observable implies a considerable increase of information, to the regime $n_{\max} > n_{\max}^{opt}$, where instead adding new observables only marginally increase the total information. $n_{\max}^{opt}$ is determined by the condition $\partial S/\partial n_{\max} = \partial \log V(n_{\max})/\partial n_{\max}$, which means that the solution is the crossing point between red and blue line in Fig. 6(c). Varying the prior, the blue line moves up and down and this moves $n_{\max}^{opt}$ by an amount $\Delta n_{\max}$. As we can see from the figure, the faster the change of slope of entropy, the smaller the range $\Delta n_{\max}$. Then, typically, big changes in the prior leads to little change in $n_{\max}^{opt}$.

V. CONCLUSIONS

In this paper, using the ME approach, we have provided rather direct evidence that the effective alignment interaction between starlings within real flocks is short-range. This result is interesting for two reasons.

First, from a biological perspective, we believe this is the first time that a short-range interaction is found without being an a priori ingredient of the model used to fit the data. In general, it is difficult to formulate a model where a qualitative crossover from short to long-range interaction occurs by tuning a parameter. Hence, what is normally done is that a certain, fixed, functional form is assumed, and its parameters fitted. Here, on the other hand, we assumed no a priori functional form of the interaction, so that the final result is completely ruled by the experimental data. We believe that short-range interaction (at least in starling flocks) can now be considered as a rather well established fact.

Secondly, our result is relevant for the maximum entropy method itself, which is increasingly used in biological inference [30–39]. A common objection to the ME method is that it is just another kind of model fitting procedure, so that, ultimately, one is prone to obtain as a *qualitative* output of the method what one feeds into the method.

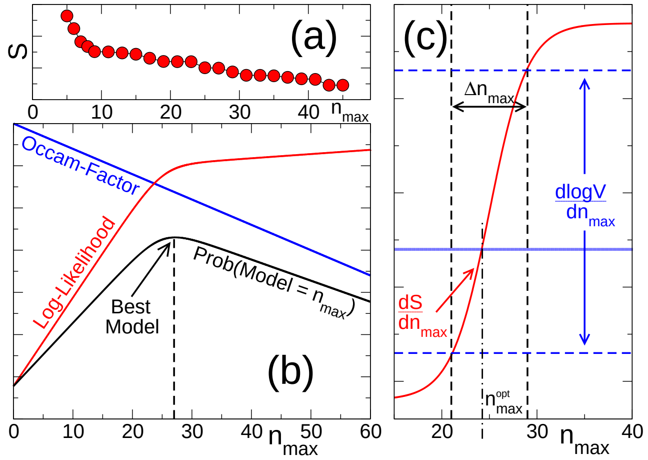

FIG. 6. (a): Entropy as a functions of number of parameters for one real event. (b): Contributions of log-likelihood and Occam factor to probability of model $P(n_{\max}|D)$. (c): Red line is derivative of entropy and each horizontal blue lines represents derivative of logarithm of Occam factor for different values of prior.

We believe that what we have found here proves otherwise. The difference between long-range and short-range interaction *is* a qualitative one, with deep consequences on the physics of the system. Yet we have seen that long-range correlation is turned into short-range interaction by the ME method, with the entropy pointing out what is the minimal number of parameters that need to be switched on, given the data. This suggests that the maximum entropy method manages to extract information from a data set with minimal bias.

We thank William Bialek for many discussions and suggestions. This work was supported by grants IIT–Seed Artswarm, ERC–StG n.257126 and US-AFOSR - FA95501010250 (through the University of Maryland).

Appendix A: Partition function in spin-wave approximation:

In order to calculate the log-likelihood we need to compute the partition function Z (Eq. (13)). In general, the exact analytical calculation of the partition functions is very hard. In the case of flocks, however, this can be done thanks to the spin-wave approximation [26]. Flocks are very ordered, with magnetization (i.e. polarization) close to 1, we can therefore expand the Hamiltonian in the small fluctuations around the mean direction of motion. The partition function can be written as,

$$Z = \int D\vec{s} \left[\prod_i \delta(|\vec{s_i}| - 1) \right] \exp \left[\sum_{i,j} J_{ij} \vec{s_i} \cdot \vec{s_j} \right] , \tag{A1}$$

where $D\vec{s} = \prod_i d\vec{s_i}$, and the $\delta-$function is enforcing the constraint that each spin has unit length. We define the global order parameter, $\vec{V} = \sum_i \vec{s_i}/N = \Phi\,\hat{n}$, where $\hat{n}$ is the unit vector and $\Phi = |\vec{V}|$ is the polarization of the flock. Each spin $\vec{s_i}$ can be rewritten in terms of the global orientation direction $\hat{n}$ and a perpendicular component to $\hat{n}$, that is $\vec{s_i} = s_i^L \hat{n} + \vec{\pi}_i$. By construction, they satisfy the following relations:

$$\vec{\pi}_i \cdot \hat{n} = 0 \ , \qquad \frac{1}{N} \sum_i s_i^L = \Phi \ , \qquad \sum_i \vec{\pi}_i = 0 \ . \tag{A2}$$

The partition function can be rewritten as,

$$Z = \int Ds^L D\vec{\pi} \left[\prod_i \delta \left(\sqrt{(s_i^L)^2 + |\vec{\pi}_i|^2 - 1} \right) \right] \delta \left(\sum_i \vec{\pi}_i \right) \exp \left[\sum_{i,j} J_{ij} (s_i^L s_j^L + \vec{\pi}_i \cdot \vec{\pi}_j) \right] , \tag{A3}$$

where $Ds^L = \prod_i ds_i^L$ and $D\vec{\pi} = \prod_i d\vec{\pi}_i$. The delta functions are taking care of the constraint on the length of each spin and of the global constraint on the $\vec{\pi}_i$. For strongly ordered flocks, $\Phi \simeq 1$ and $|\vec{\pi}_i| \ll 1$. Then, at second order, $s_i^L \simeq 1 - |\vec{\pi}_i|^2/2$. Performing the integral over s^L, the partition function becomes,

$$Z = \int D\vec{\pi} \left[\prod_i \frac{1}{\sqrt{1 - |\vec{\pi}_i|^2}} \right] \delta \left(\sum_i \vec{\pi}_i \right) \exp \left[- \sum_{i,j} A_{ij} \vec{\pi}_i \cdot \vec{\pi}_j + \sum_{i,j} J_{ij} \right] , \tag{A4}$$

where

$$A_{ij} = \delta_{ij} \left(\sum_k J_{ik} \right) - (1 - \delta_{ij}) J_{ij} . \tag{A5}$$

For strongly ordered flocks, the product $\prod_i 1/\sqrt{1 - |\vec{\pi}_i|^2}$ can be neglected (we have explicitly checked that the corrections due to this term are indeed negligible). Therefore, we can write

$$Z = \int D\vec{\pi}\ \delta \left(\sum_i \vec{\pi}_i \right)\ \exp \left[- \sum_{i,j} A_{ij} \vec{\pi}_i \cdot \vec{\pi}_j + \sum_{i,j} J_{ij} \right] . \tag{A6}$$

Since $J_{ij} = J_{ji}$ (and then $A_{ij} = A_{ji}$) we benefit from the spectral theorem for symmetric matrices. The matrix A_{ij} is diagonalizable, its eigenvalues are real and its eigenvectors form an orthonormal basis. Moreover, the condition $\sum_j A_{ij} = 0$ means that the matrix A_{ij} is a positive semidefinite matrix having the smallest eigenvalue $a_1 = 0$, and all other eigenvalues positive. Let a_k be the eigenvalue corresponding to the eigenvector $\mathbf{w}_k$. The eigenvector $\mathbf{w}_k$ satisfies the usual relation,

$$\sum_j A_{ij} w_j^k = a_k w_i^k. \tag{A7}$$

It can be easily seen that the eigenvector $\mathbf{w^1}$ corresponding to a_1 is constant and it is given by $(1/\sqrt{N}, 1/\sqrt{N}, .., 1/\sqrt{N})$. We can rewrite the integral in the orthonormal basis defined by $\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^N$:

$$Z = \int D\vec{\pi}' \ \delta(\vec{\pi}'_1) \exp\left[-\sum_{k=1}^{N} a_k |\vec{\pi}'_k|^2 + \sum_{i,j} J_{ij} \right], \tag{A8}$$

where $\vec{\pi}'_k = \sum_i w_i^k \vec{\pi}_i$. From this form it is clear that the role of the δ-function over the $\vec{\pi}'_1$ is exactly to eliminate the zero mode from the integral. Performing the Gaussian integral in two dimensions, we obtain,

$$\log Z = -\sum_{k>1} \log a_k + \sum_{i,j} J_{ij}, \tag{A9}$$

where the irrelevant constant terms have been neglected.

Appendix B: Coarse-graining

The experimental observable we consider in flocks is the correlation function $\hat{C}(n)$. In principle, this correlation function can be computed for each value of the topological distance n but, as discussed in the paper, it is safe to consider only $\hat{C}(n)$ up to $n = n_{\max} \ll N$. Furthermore, in order to decrease the number of parameters and speed up the numerical task, we can consider a "coarse graining", that is a binning of n with generic increment $\Delta n \geq 1$. This means that we include in the same observable $\hat{C}(n)$ contributions from the distances $n, n+1, ..., n+\Delta n - 1$. Hence, we have,

$$\hat{C}(n) = \frac{\sum_{i,j} \vec{s_i} \cdot \vec{s_j} \ \delta(k_{ij} - n)}{\sum_{i,j} \delta(k_{ij} - n)}, \tag{B1}$$

where $\{\vec{s_i}\}$ are unit vectors and $k_{ij} = n$ if j is the n^{th} neighbor of i. $\delta(k-n)$ is a "modified" Kronecker's δ that takes into account the binning of n,

$$\delta(k-n) = \begin{cases} 1 & \text{if} \quad n \leq k < n + \Delta n, \\ 0 & \text{otherwise.} \end{cases} \tag{B2}$$

Note that for $\Delta n = 1$ the model reduces to the one described in the main text.

For each observable $\hat{C}(n)$ the associated Lagrange multiplier is denoted by λ_n. The maximum entropic Hamiltonian consistent with these observables is

$$H = \sum_n{}' \lambda_n \hat{C}(n), \tag{B3}$$

where the symbol $\sum'$ means that the sum stops at $n_{\max}$ and that we sum only over the "bins" of the coarse graining, that is $n = 1, 1+\Delta n, 1+2\Delta n, ..., n_{\max}$. Physically, λ_n is the *total* interaction strength for bin n for a flock (extensive), whereas the interaction strength $J(n)$ defined in the main text is the strength for an *individual pair* within bin n (intensive) and hence $J(n) = -\lambda_n/(N\Delta n)$.

Using coarse-grained variables does not change formally the computation, as outlined in the main text. Indeed if we define $\hat{J}(k_{ij})$ such that

$$\hat{J}(k_{ij}) = \sum_n{}' J(n) \delta(k_{ij} - n), \tag{B4}$$

(note that with $\Delta n = 1$, $\hat{J}(k_{ij}) \equiv J(k_{ij})$), we can easily verify that Eq. (B3) reads as a classical Heisenberg model

$$H(\{\vec{s_i}\}) \ = \ -\sum_{i,j} \hat{J}(k_{ij}) \ \vec{s_i} \cdot \vec{s_j} \equiv \ -\sum_{i,j} J_{ij} \vec{s_i} \cdot \vec{s_j}. \tag{B5}$$

where the symmetrized interactions matrix J_{ij} are given by

$$J_{ij} \equiv \frac{1}{2}[\hat{J}(k_{ij})+\hat{J}(k_{ji})] = \frac{1}{2}\sum_n{}' J(n)[\delta(k_{ij}-n)+\delta(k_{ji}-n)], \tag{B6}$$

The partition function can be computed using the spin-wave expansion, as described in Appendix A, where, now, the eigenvalues a_k refer to the matrix J_{ij} in Eq. (B6). The log-likelihhood takes the form

$$\log\mathcal{L} = -\log Z - \lambda_n\langle\hat{C}(n)\rangle_{\rm expt} = -\log Z + N\Delta n\sum_n{}' J(n)\langle\hat{C}(n)\rangle_{\rm expt} \tag{B7}$$

The coarse graining procedure can be easily generalized to the anisotropic case. We consider the transverse and longitudinal coarse grained correlations (see Eqs. (27))

$$\hat{C}^L(n) = \frac{\sum_{i,j}\vec{s_i}\cdot\vec{s_j}\delta(k_{ij}-n)\Theta\left(|\cos(\theta_{ij})|-1/2\right)}{\sum_{i,j}\delta(k_{ij}-n)\Theta\left(|\cos(\theta_{ij})|-1/2\right)}, \tag{B8}$$

$$\hat{C}^T(n) = \frac{\sum_{i,j}\vec{s_i}\cdot\vec{s_j}\delta(k_{ij}-n)\Theta\left(1/2-|\cos(\theta_{ij})|\right)}{\sum_{i,j}\delta(k_{ij}-n)\Theta\left(1/2-|\cos(\theta_{ij})|\right)}, \tag{B9}$$

where, we remind, θ_{ij} is the angle formed by $\vec{r}_{ij}=(\vec{r}_j-\vec{r}_i)$ and the flock's direction of motion $\vec{V}=1/N\sum_i\vec{s_i}$. $\Theta(x)$ is the Heaviside step function and the factor 1/2 divides the space evenly between the two sectors. The δ-function bears the same meaning as in (Eq. (B2)) and identifies pairs belonging to the same bin centered around the topological distance n and of width Δn.

Following the same method as above, when using coarse grained correlations we need to introduce different Lagrange multipliers $\lambda_n^{L,T}$ for each bin (rather than for each discrete value of n). The ME Hamiltonian then reads

$$H = \sum_n{}' \lambda_n^L\hat{C}^L(n) + \lambda_n^T\hat{C}^T(n) \tag{B10}$$

Also this Hamiltonian can be written as an Heisenberg-like Hamiltonian. The procedure is slightly more complicated than in the isotropic case.

We first define the fraction of neighbors that lie in the longitudinal and transversal sector for each bin around n,

$$p^L(n) = \tfrac{1}{N\Delta n}\sum_{i,j}\delta(k_{ij}-n)\Theta\left(|\cos(\theta_{ij})|-1/2\right), \tag{B11}$$

$$p^T(n) = \tfrac{1}{N\Delta n}\sum_{i,j}\delta(k_{ij}-n)\Theta\left(1/2-|\cos(\theta_{ij})|\right), \tag{B12}$$

$$\tag{B13}$$

These quantities of course satisfy the relation $p^L(n)+p^T(n)=1$. At this point, we can express the Lagrange multipliers $\lambda_m^{L,T}$ (defined for a given bin) in terms of the effective longitudinal and transversal interactions $J^{L,T}(n)$ (defined for each pair of individuals at distance n). We have $J^{L,T}(n) = -\lambda_n^{L,T}/(p^{L,T}(n)N\Delta n)$. Then, as above, we introduce the pairwise interactions $\hat{J}^{L,T}(k_{ij})$

$$\hat{J}^{L,T}(k_{ij}) = \sum_n{}' J^{L,T}(n)\delta(k_{ij}-n)\Theta_{ij}^{L,T}, \tag{B14}$$

where $\Theta_{ij}^{L,T}\equiv\Theta(\pm|\cos(\theta_{ij})|\mp 1)$ and selects pairs that contribute to, respectively, the longitudinal and the transverse sectors. With these substitutions the Hamiltonian acquires an Heisenberg form,

$$H(\{\vec{s_i}\}) = -\sum_{i,j}\left[\hat{J}^L(k_{ij})+\hat{J}^T(k_{ij})\right]\vec{s_i}\cdot\vec{s_j} \equiv -\sum_{i,j}J_{ij}\vec{s_i}\cdot\vec{s_j} \tag{B15}$$

where J_{ij} is now the symmetric part of the matrix $\hat{J}^L(k_{ij})+\hat{J}^T(k_{ij})$

The log-likelihhood then becomes

$$\begin{aligned}\log\mathcal{L} &= -\log Z - \lambda_n^T\langle\hat{C}^T(n)\rangle_{\rm expt} - \lambda_n^L\langle\hat{C}^L(n)\rangle_{\rm expt}\\ &= -\log Z + N\Delta n\sum_n{}'\left[p^T(n)J^T(n)\langle\hat{C}^T(n)\rangle_{\rm expt}) + p^L(n)J^L(n)\langle\hat{C}^L(n)\rangle_{\rm expt})\right].\end{aligned} \tag{B16}$$

All figures displayed in this paper are obtained using a coarse graining with $\Delta n = 2$ for $n = 2, \cdots n_{\rm max}$, and $\Delta n = 1$ for $n = 1$. We also used $\Delta n = 1$, results are fully consistent with the larger coarse graining, just more noisy.

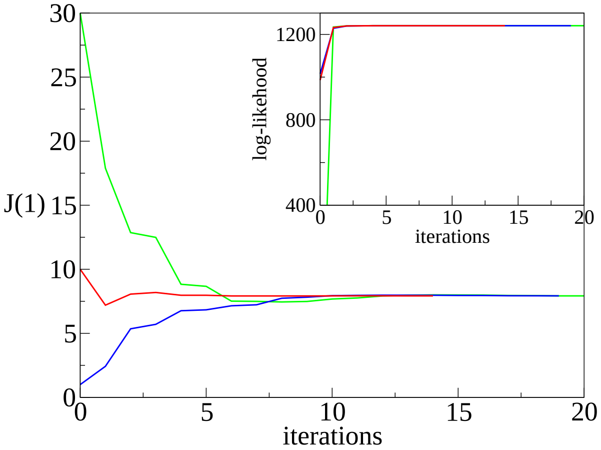

FIG. 7. Stability and convergence of the numerical method: three very different initial guesses lead to the same value of $J(1)$ (this is generically true for all other $J(n)$). Inset: also the log-likelihood reaches the same asymptotic value with different initial conditions. This implies that it exists a stable and unique global maximum of the log-likelihood.

Appendix C: Computing the derivatives of the Log-Likelihood

Using the analytical expression of the partition function Eq. (A9), we can write the expressions of the log-likelihood (Eq. (8)) for the case of full interaction,

$$\log \mathcal{L} = -\log Z + N\Delta n {\sum_n}' J(n)\langle \hat{C}(n)\rangle_{\rm expt} = \sum_{k>1} \log a_k - N\Delta n {\sum_n}' J(n)(1-\langle \hat{C}(n)\rangle_{\rm expt}). \tag{C1}$$

Similarly, the log-likelihood function for the anisotropic case is given by,

$$\log \mathcal{L} = \sum_{k>1} \log a_k - N\Delta n {\sum_n}' p^T(n) J^T(n)(1-\langle \hat{C}^T(n)\rangle_{\rm expt}) - N\Delta n {\sum_n}' p^L(n) J^L(n)(1-\langle \hat{C}^L(n)\rangle_{\rm expt}). \tag{C2}$$

The condition for maximizing the log-likelihood is $\partial \log \mathcal{L}/\partial J(n) = 0$. Let us consider, for the moment, the isotropic case. The derivative of the second term of the log-likelihood with respect to $J(n)$ is trivial. However, differentiating the partition function is far less trivial, as the eigenvalues a_k are very complicated functions of the $\{J(n)\}$. We will calculate $\partial a_k/\partial J(n)$ by using perturbation theory. Suppose that we perturb $J(n)$ by some small amount, $J(n) \to J(n) + \epsilon$, where ϵ is infinitesimal. The perturbation makes A_{ij} change into,

$$\widetilde{A}_{ij}(\epsilon) = A_{ij} + \epsilon \gamma_{ij}(n), \tag{C3}$$

where we introduced a symmetric matrix

$$\gamma_{ij}(n) = \frac{\partial A_{ij}}{\partial J(n)} = \sum_{l,m} \frac{\partial A_{ij}}{\partial J_{lm}} \frac{\partial J_{lm}}{\partial J(n)}. \tag{C4}$$

Due to this small perturbation the eigenvalue a_k and its eigenvector $\mathbf{w}^k$ change by small amount,

$$\widetilde{a}_k(\epsilon) = a_k + \epsilon \xi_k + \mathcal{O}(\epsilon^2), \tag{C5}$$

$$\widetilde{w}_i^k(\epsilon) = w_i^k + \epsilon g_i^k + \mathcal{O}(\epsilon^2). \tag{C6}$$

For the $\widetilde{A}(\epsilon)$ matrix we can write,

$$\sum_j \widetilde{A}_{ij}(\epsilon)\widetilde{w}_j^k(\epsilon) = \widetilde{a}_k(\epsilon)\widetilde{w}_i^k(\epsilon). \tag{C7}$$

Through some algebra it is quite straightforward to show that,

$$\widetilde{a}_k(\epsilon) = a_k + \epsilon \sum_{ij} \gamma_{ij}(n) w_i^k w_j^k + \mathcal{O}(\epsilon^2). \tag{C8}$$

Therefore, the derivative of the eigenvalue a_k can be written as

$$\frac{\partial a_k}{\partial J(n)} = \lim_{\epsilon \to 0} \frac{\widetilde{a}_k(\epsilon) - a_k}{\epsilon} = \sum_{i,j} \gamma_{ij}(n) w_i^k w_j^k . \tag{C9}$$

To obtain the form of matrix $\gamma_{ij}(n)$ we use first Eq. (A5) and Eq. (B6) from which

$$\begin{aligned} \frac{\partial A_{ij}}{\partial J_{lm}} &= \delta_{il}\left(\delta_{ij} - (1-\delta_{ij})\delta_{jm}\right) \\ \frac{\partial J_{lm}}{\partial J(n)} &= \frac{1}{2}[\delta(k_{lm} - n) + \delta(k_{ml} - n)] \end{aligned} \tag{C10}$$

then from Eq. (C4)

$$\gamma_{ij}(n) = \frac{1}{2}\delta_{ij}\left[\sum_m (\delta(k_{im} - n) + \delta(k_{mi} - n))\right] - \frac{1}{2}(1-\delta_{ij})(\delta(k_{ij} - n) + \delta(k_{ji} - n)). \tag{C11}$$

In the same way, using the expression for the anisotropic J_{ij} we obtain $\gamma_{ij}^{L,T}(n)$ for the anisotropic case,

$$\gamma_{ij}^{L,T}(n) = \frac{1}{2}\delta_{ij}\left[\sum_m (\delta(k_{im} - n) + \delta(k_{mi} - n))\Theta_{im}^{L,T}\right] - \frac{1}{2}(1-\delta_{ij})(\delta(k_{ij} - n) + \delta(k_{ji} - n))\Theta_{ij}^{L,T}. \tag{C12}$$

Now, using Eq. (C9), it becomes easy to calculate the derivatives of the log-likelihood (Eq. (C1)) w.r.t each of its variable $J(n)$. Imposing its maximization we obtain,

$$1 - \langle \hat{C}(n) \rangle_{\text{expt}} = \frac{1}{N\Delta n} \sum_{k>1} \frac{1}{a_k} \frac{\partial a_k}{\partial J(n)} = \frac{\text{Tr}[A^{-1}\gamma(n)]}{N\Delta n}. \tag{C13}$$

Similarly, for the anisotropic case the maximization of Eq. (C2) gives,

$$1 - \langle \hat{C}^{L,T}(n) \rangle_{\text{expt}} = \frac{1}{p^{L,T}(n) N\Delta n} \sum_{k>1} \frac{1}{a_k} \frac{\partial a_k}{\partial J^{L,T}(n)} = \frac{\text{Tr}[A^{-1}\gamma^{L,T}(n)]}{p^{L,T}(n) N\Delta n}. \tag{C14}$$

Appendix D: Numerical maximization of the log-likelihood

The analytical expressions of the partition functions and its derivatives are not enough to analytically optimize the log-likelihood (Eq. (16), Eq. (C2)). The reason is the following: the partition function and its derivatives are functions of the eigenvalues and eigenvectors of the network matrix A_{ij} and it is not possible to diagonalize such $N \times N$ matrix (N is the number of birds of the flock) without the help of any numerical method. However, as we discuss in this section, knowing explicitly the derivatives enormously simplifies the numerical procedure.

Finding numerically the optimum of a multidimensional function is always tricky. Two practical issues must be addressed. First, the solution must be stable, i.e., different initial guesses should lead to the same solution. Second, the computation should be numerically efficient. There are two ways to approach such problem: (i) without providing the analytical expressions of the derivatives; (ii) providing the analytical expressions of the derivatives. In the first case the number of iterations needed for the optimization is much larger than in the second case, as we provide less information. The reason is the following. Given an initial guess (i.e. a set of $\{J(n)\}$) the numerical optimization algorithm must explore the space of the parameters $J(n)$ around the initial values to find the direction leading to the maximum. This practically means computing the derivatives of the log-likelihood. To compute numerically such derivatives the algorithm must evaluate the log-likelihood not only at the starting point, but also for for small increments of the $\{J(n)\}$. Since the log-likelihood depends on the eigenvalues of A_{ij}, this in turn implies that this matrix must be diagonalized more than once. On the contrary, if the analytical expressions of the derivatives are provided, only the eigenvalues and the eigenvectors of A_{ij} at the starting point are required (see Appendix C) and A must be diagonalized only once. During each optimization step the most time-consuming part is precisely the diagonalization of the matrix A_{ij}. Therefore the optimization time is significantly smaller with method (ii) than with (i), and the whole computation gets much more efficient. Furthermore, as the dimension of the log-likelihood function gets larger, the number of iterations in the first case increases very rapidly. Finally, among all optimization algorithms currently available the most performing ones (in terms of robustness and speed) are the ones that use the

EVENT	N	Φ	L (m)	n_c^{exp}	n_c^{step}	r_c^{exp}(m)
21–06	717	0.973	32.1	7.41	11.73	2.00
25–10	1047	0.991	33.5	9.56	14.30	1.93
25–11	1176	0.959	43.3	12.01	15.03	1.99
28–10	1246	0.982	36.5	4.92	10.21	1.27
29–03	440	0.963	37.1	4.46	7.67	1.94
31–01	2126	0.844	76.8	6.11	12.37	2.97
32–06	809	0.981	22.2	7.43	12.50	1.39
42–03	431	0.979	29.9	7.79	14.60	2.08
49–05	797	0.995	19.2	6.18	11.25	1.24
57–03	3242	0.978	85.7	8.51	14.19	2.67
58–06	442	0.984	23.1	7.39	12.89	1.63
58–07	554	0.977	19.1	7.23	13.79	1.63
63–05	890	0.978	52.9	5.26	10.21	1.98
69–09	239	0.985	17.1	10.56	16.91	1.92
69–10	1129	0.987	47.3	9.11	15.30	2.39
69–19	803	0.975	26.4	14.76	21.56	1.97
72–02	122	0.992	10.6	8.62	11.37	1.32
77–07	186	0.978	9.1	5.97	12.36	1.17
20111125-2	505	0.972	34.4	11.31	16.36	1.84
20111214-4-1	139	0.985	32.8	5.24	9.97	1.64
20111214-4-2	156	0.983	31.5	8.13	10.23	2.52
20111215-1	394	0.994	49.8	8.48	20.62	1.68

TABLE I. **Flocks Data**: Each line represents a different flocking event. N is the number of individuals in the flock, Φ the average polarization, L the size of the flock (maximum distance between two birds), n_c^{exp} the exponential decay range computed in this work and n_c^{step} the interaction range of the step model of [26]. Finally, r_c^{exp} represents the typical metric distance corresponding to the topological distance n_c^{exp}: it can be seen that it is always much smaller than the flock's size.

analytical derivatives. Practically speaking, for the largest flocks method (ii) is more than 10 times faster than (i). Therefore, the analytical expressions of the derivatives that we have (painfully) worked out in the previous sections are very useful to obtain a stable numerical solution in an efficient way.

We use the minimizing routine *gsl multimin fminimizer nmsimplex2*, belonging to the gnu scientific library [56]. This optimization algorithm is based on Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [57]. We provide as an input of the routines the analytical expressions of Z and of the derivatives of Z with respect to $J(n)$ (or to $J^L(n)$ and $J^T(n)$). For the diagonalization, we use the *gsl eigen symmv* belonging to the gnu scientific library [56].

In Fig. 7, we plot the behavior of the parameter $J(1)$ and of the log-likelihood vs the iteration time, for three different initial conditions of $J(1)$. It is clear that the solution is very stable and it is also reached very quickly.

Appendix E: Data set

Experimental data were obtained from field observations on large flocks of starlings, (*Sturnus vulgaris*), in the field. Three dimensional trajectories of positions and velocities of each bird are obtained using stereometric photography and computer vision techniques [4, 7, 25, 28, 47–49]. As summarized in Table I, we have analyzed 22 distinct flocking events, with sizes ranging from 122 to 3242 individuals and linear extensions from 9.1 to 85.7 m. All these events belong to two different sets. The first set (events from 21-06 to 77-07 in Table I) was taken in the period 2005-2008, with cameras shooting at 10 frames-per-second (fps). [25, 28]. The second set (last 4 events in Table I) was collected in the period between 2010-2012, with cameras shooting at 170fps [7, 49]. All the events correspond to strongly ordered flocks, with polarization between $\Phi = 0.844$ and $\Phi = 0.995$, hence justifying the spin wave expansion. The duration of the observed events is on average 6 seconds and it ranges between 2.8 and 11.6 seconds. The number of frames varies between 14 and 58 frames per event (with mean 30). These scales are set by experimental constraints. In a stereoscopic experiment a flocking event is filmed by several machine vision cameras located at different positions.

To reconstruct the individual 3D trajectories the flock must be in the common field of view of all the cameras: given the flock's typical distance from the apparatus (100-300 m), after 10-12 seconds at most the flock is out of the field of view (this time being shorter the larger/closer is the flock). Besides, the amount of digital information per second that can be grabbed by a high resolution stereo set-up is limited, which also sets a constraint on the amount of consecutive digital images that can be retrieved at high frequency. We note that these time durations represent significant scales in terms of the collective motion of natural flocks: starlings fly at approximately 10 m/s and a flock of thousands birds can perform a collective turn (global change of direction) in just a few seconds [7].

[1] S. Camazine *et al.*, *Self-organization in Biological Systems* (Princeton University Press, Princeton, 2003).
[2] I. D. Couzin and J. Krause, Adv. Study. Behav **32**, 1 (2003).
[3] Vicsek T., Zafeiris A. Phys. Rep. **517**, 71 (2012).
[4] A. Cavagna *et al.*, Proc. Natl. Acad. Sci. **107**, 11865 (2010).
[5] A. Attanasi *et al.*, PLoS computational biology, 10(7), e1003697. Physical review letters 113.23 (2014): 238102.
[6] D. J. T. Sumpter, J. Buhl, D. Biro, and I. D. Couzin, Theor. Biosci. **127**, 177 (2008).
[7] A. Attanasi *et al.*, Nature physics **10**, 691-696 (2014).
[8] J. Herbert-Read, J. Buhl, F. Hu, A. Ward & D.J. Sumpter, arXiv preprint arXiv:1409.6750 (2014).
[9] M.C. Marchetti, et al. Reviews of Modern Physics **85**, 1143 (2013)
[10] S. Ramaswamy, Annu. Rev. Condens. Matter Phys. **1**, 323 (2010)
[11] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Phys. Rev. Lett. **75**, 1226 (1995).
[12] G. Grégoire, H. Chaté, Phys. Rev. Lett. **92**, 025702 (2004)
[13] H. Chaté, et al. Eur. Phys. J. B **64**, 451 (2008)
[14] F. Ginelli, H. Chaté, Phys. Rev. Lett. **105**, 168103 (2010)
[15] C. K. Hemelrijk and H. Hildenbrandt, Interface focus **2**, 726 (2012).
[16] J. Toner and Y. Tu, Phys. Rev. Lett. **75**, 4326 (1995).
[17] J. Toner and Y. Tu, Phys. Rev. E **58**, 4828 (1998).
[18] D. J. T. Sumpter, *Collective Animal Behavior* (Princeton University Press, New Jersey, 2010).
[19] D. J. G. Pearce, A. M. Miller, G. Rowlands, and M. S. Turner, Proc. Natl. Acad. Sci. published ahead of print July 7, 2014, doi:10.1073/pnas.1402202111 (2014).
[20] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, Proc. Natl. Acad. Sci. (2010).
[21] J. Gautrais *et al.*, PLoS Comput. Biol **8**, e1002678 (2012).
[22] Y. Katz, K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin, Proc. Natl. Acad. Sci. **108**, 18720 (2011).
[23] J. G. Puckett, D. H. Kelley, and N. T. Ouellette, Scientific reports **4**, 4766 (2014).
[24] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[25] M. Ballerini *et al.*, Proc. Natl. Acad. Sci. **105**, 1232 (2008).
[26] W. Bialek *et al.*, Proc. Natl. Acad. Sci. **109**, 4786 (2012).
[27] W. Bialek *et al.*, Proc. Natl. Acad. Sci. **111**, 7212 (2014).
[28] M. Ballerini *et al.*, Animal behavior **76**, 201 (2008).
[29] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003).
[30] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[31] J. Shlens *et al.*, J. Neurosci. **29**, 8254 (2006).
[32] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, Proc. Natl. Acad. Sci. **103**, 19033 (2006).
[33] A. Tang *et al.*, J. Neurosci. **28**, 505 (2008).
[34] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. **106**, 67 (2009).
[35] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
[36] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc. Natl. Acad. Sci. **107**, 5405 (2010).
[37] G. Tkačik, ArXiv preprint , 1006.4291 (2010).
[38] G. Tkačik *et al.*, J. Stat. Mech. **2013**, P03011 (2013).
[39] G. Tkačik *et al.*, PLoS Comput. Biol. **10**, e1003408 (2014).
[40] A. Cavagna et al., Phys. Rev. E **89**, 042707 (2014)
[41] M. Castellana, et al. ArXiv preprint arXiv:1412.8654 (2014).
[42] Y. Roudi and J. Hertz, Phys. Rev. Lett. **106**, 048702 (2011).
[43] E. Van der Straeten. Maximum Entropy Estimation of Transition Probabilities of Reversible Markov Chains. Entropy **11**, 867887 (2009)
[44] O. Marre, et al. Phys. Rev. Lett. **102**, 138101 (2009).
[45] C. E. Shannon, Bell Sys. Tech. **27**, 379 (1948).
[46] Note that the topological distance, n, is a discrete variable, hence the correlation function $\hat{C}(n)$ is in fact a discrete set of values and it should be indicated as $\hat{C}_n$. However, we prefer the $\hat{C}(n)$ notation, as it is reminiscent of the standard metric definition of correlation function, $\hat{C}(r)$. For the same reason we inappropriately use Dirac's notation $\delta(k_{ij} - n)$ for what is actually a Kronecker's $\delta_{k_{ij},n}$.

[47] A. Cavagna *et al.*, Animal behavior **76**, 217 (2008).

[48] A. Cavagna, I. Giardina, A. Orlandi, G. Parisi, and A. Procaccini, Animal behavior **76**, 237 (2008).

[49] A. Attanasi *et al.*, ArXiv preprint , 1305.1495 (2013).

[50] Indeed the same calculation performed in Sec. II can be repeated in a metric fashion (using the metric distance, r, rather than the topological distance, n); in this case, we find that the metric range r_c scales with the density, thus confirming the result of [25] that the interaction in starling flocks is based on topological distance.

[51] A. Cavagna , et al. Proc. R. Soc. B **280**, 20122484 (2013)

[52] T. M. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[53] F. Raynaud, PhD Thesis, Université Paris Diderot, 2009.

[54] M. Castellana, W. Bialek, A. Cavagna, I. Giardina, arXiv preprint arXiv:1412.8654 (2014).

[55] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw-Hill, New York, 1978).

[56] B. Gough, *GNU Scientific Library Reference Manual - Third Edition*, 3rd ed. (Network Theory Ltd., 2009).

[57] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. (Wiley, 1987).