Short-range interaction vs long-range correlation in bird flocks

Andrea Cavagna\textsuperscript{a,b,c}, Lorenzo Del Castello\textsuperscript{a,b}, Supravat Dev\textsuperscript{a,b}, Irene Giardina\textsuperscript{a,b,c}, Stefania Melillo\textsuperscript{a,b}, Leonardo Parisi\textsuperscript{a,b,d}, and Massimiliano Viale\textsuperscript{a,b}

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

We use the maximum entropy method to study how the strength of effective alignment between birds depends on distance. We find in all analyzed flocks that the interaction decays exponentially. Such short-range form is noteworthy, considering that the velocity correlation that is input of the calculation is long-ranged. We use our method to study the directional anisotropy in the alignment interaction and find that the interaction strength along the direction of motion is weaker than in the transverse direction, which may account for the anisotropic spatial distribution of birds observed in natural flocks.

In general, short-range interaction is sufficient to produce spontaneous symmetry breaking and system-level coordination. Indeed, there now seems to be some consensus in the field of collective animal behaviour that interactions are short-ranged \cite{12}. 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 correlation than some arcane physical theorem, not to mention criticality. For example, birds’ vision is likely to span the entire size of a flock and it has been recently proposed that a vision-based long-range interaction is at the basis of flocking behaviour \cite{12}.

Although the access to large scale empirical data on animal groups in the last decade or so has considerably advanced our understanding of the problem, we still have no direct and unbiased proof of whether interaction is short- or long-ranged. Very significant results have been obtained by fitting biological models to the data \cite{12}. Yet the problem with model fitting is that it may be tricky to distinguish the bona fide properties of the system under investigation from the a priori ingredients of the model used to fit the data. Alternatively, interaction has been assessed by using some structural proxy of it; for example, in \cite{9} it has been found experimentally that the anisotropy in the nearest neighbours spatial distribution decays in a short-range way. However, through this kind of structural proxies one does not attain direct access to the interaction.

A different approach is to use the maximum entropy method, which is different from standard model fitting in that the model is dictated by the available experimental observables \cite{20}. For bird flocks this was done in \cite{10}. However, even in that case the interaction was assumed to be step-like, meaning that each bird was interacting with equal strength with all its first $n_c$ neighbours and not interacting at all beyond $n_c$. Because of the step-like assumption, the calculation of \cite{10} provided no information on the way the interaction decays with distance. Moreover, the only evidence of the interaction being short-range was that the inferred $n_c$ did not scale with the size of the system, which rather indirect. Therefore, a full characterization of the interaction as a function of distance in animal groups and a direct assessment of its range are still lacking.

Here we perform a novel maximum entropy calculation able to derive the dependence on distance of the effective alignment interaction in real flocks of birds. We find that the strength $J$ of the alignment decays exponentially with the topological distance $n$ between birds \cite{9}, hence proving that the interaction is short-ranged. The input of our calculation is the velocity correlation function of distance in animal groups and a direct assessment of its range are still lacking.

The first step of the maximum entropy (ME) method is to identify the set of variables that defines the mi-
crostate of the system under investigation. We are interested in the interaction that is responsible for the alignment of the directions of motion of the birds. Hence we define 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 an observable, 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, $\langle \vec{s} \rangle, \langle \vec{s} \cdot \vec{s} \rangle, \ldots$. Each one of these $m$-points correlations will generate $m$-points interaction terms and the most economic prescription is 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 [22]) the use of pairwise interaction ($m = 2$) allows to accurately predict the 4-points correlations [10, 11]. Therefore, our experimental observable will be the two-point correlation function,

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

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 [2], this correlation is function of the topological distance (i.e. order of neighborhood), $n$, rather than of the metric distance, $r$ [2], so that $k_{ij}$ 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 (thus non-symmetric). This implies $\sum_{ij} \delta(k_{ij} - n) = N$, and this is why the normalization in (1) is easier than in its metric counterpart [3] [16].

The ME method consists in finding the probability distribution $P(\{\vec{s}_i\})$ that maximizes the entropy [20],

$$S[P] = -\int D\vec{s} \, P(\{\vec{s}_i\}) \log P(\{\vec{s}_i\}),$$  (2)

under the constraint that the distribution reproduces the experimental observables [11],

$$\langle \hat{C}(n; \{\vec{s}_i\}) \rangle_{\text{exp}} = \langle \hat{C}(n; \{\vec{s}_i\}) \rangle_P .$$  (3)

This constrained maximization is achieved by introducing one Lagrange multiplier, $J(n)$, for each experimental quantity that we are fixing, $\hat{C}(n)$. As it always happens within ME, the distribution obtained in this way has the form of an exponential of the product of the Lagrange multipliers times the observables (see Supplemental Material-SM),

$$P(\{\vec{s}_i\}) = \frac{1}{Z} \exp\left\{ N \sum_n J(n) \hat{C}(n) \right\} = \frac{1}{Z} \exp\left\{ \sum_{ij} J(k_{ij}) \vec{s}_i \cdot \vec{s}_j \right\},$$  (4)

where $Z$ is the normalizing partition function. The probability distribution [11] corresponds to the Hamiltonian,

$$H = -\sum_{ij} J(k_{ij}) \vec{s}_i \cdot \vec{s}_j .$$  (5)

Therefore, the (discrete) function $J(n)$ is the strength of the effective alignment interaction between birds at topological distance $n$. In order to understand whether the effective alignment interaction in the system is short- or long-ranged we need to know how $J$ decays with $n$.

For a given experimental configuration of the positions and of the velocities, the values of the Lagrange multipliers, $J(n)$, are obtained by imposing constraint [3], namely by asking that the experimentally measured correlation function is the same as that computed from the probability distribution $P(\{\vec{s}_i\})$. This amounts to solving, for each $n$, the equation,

$$\langle \hat{C}(n; \{\vec{s}_i\}) \rangle_{\text{exp}} = \int \frac{D\vec{s}}{Z} \exp\left\{ \sum_{ij} J(k_{ij}) \vec{s}_i \cdot \vec{s}_j \right\} \hat{C}(n; \{\vec{s}_i\}) .$$  (6)

The l.h.s. of this equation is a number coming out of the experiment, while the r.h.s. is a function of the interaction couplings $J(n)$. We apply the method to flocks of European starlings (see [8, 17–19] and SM for details on the data). Starling flocks are highly ordered systems [18]. Therefore, the connected correlation function $C(n)$ compared to the interaction $J(n)$ (both normalized by their $n = 0$ value to show on the same scale).

FIG. 1. **Left** Connected correlation function $C(n)$ compared to the interaction $J(n)$ (both normalized by their $n = 0$ value to show 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 $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63$, and $n_c = 7.41$. **Bottom.** Event 21-06, $N = 717$; $J_0 = 25.63
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 \( n_{\text{max}} \) can be understood also at the level of the entropy. The value of the entropy as a function of \( n_{\text{max}} \) after maximization tells us how much information we gain \[23\] by adding more experimental data \( (\hat{C}(n)) \) and by inferring more parameters \( (J(n)) \). We can see from Fig. 2b that the entropy decays very fast up to a certain \( n_{\text{max}} \approx 15 \), and then the decay becomes slower and linear. This means that we are gaining real information for \( n_{\text{max}} \leq 15 \), but after that we are simply fitting the noise and there is no more useful information to be gained by increasing \( n_{\text{max}} \). 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 (see SM for more details).

We now address our second question: is the alignment interaction isotropic? To answer we introduce a simple generalization of the previous calculation. 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 \( \theta_{ij} \) be the angle formed by \( \vec{r}_{ij} \) 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). With this definition the two sectors have the same volume. We then define the longitudinal (L) and transverse (T) correlation functions, which are simply the average correlations in their relative sectors,

\[
\tilde{C}_{L,T}(n) = \frac{\sum_{i,j} \delta(|k_{ij} - n|) \Theta(\pm |\cos(\theta_{ij})| \mp 1/2)}{\sum_{i,j} \delta(|k_{ij} - n|) \Theta(\pm |\cos(\theta_{ij})| \mp 1/2)} .
\]

Using these new observables we can infer through the ME method the longitudinal interaction \( J^L(n) \) and the transverse interaction \( J^T(n) \), by using the same procedure as above (see SM for details). The result shows that the interaction between nearest neighbors in the transverse direction is slightly, but significantly, stronger than that in the longitudinal direction. In other words, \( J^T(n = 1) > J^L(n = 1) \), on average (Fig. 3). This anisotropic character of the interaction is very short-ranged, though, as it already disappears by the second nearest neighbor, \( J^T(n = 2) \approx J^L(n = 2) \) (Fig. 3).

This anisotropy is interesting. 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 neighbour 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 neighbour has more severe consequences (in terms of collision) than...
FIG. 3. Log-log plot of $J^L$ vs $J^T$ for $n = 1$, and $n = 2$ for all flocks. The full line is the identity.

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 [11]; it would therefore be interesting to extend the present calculation to the case of speed.

The anisotropic interaction that we find is likely to have some impact on the spatial arrangement of neighbours. It was indeed found in [9] that the neighbours of a bird are more easily found in the transverse than in the longitudinal direction. Although this anisotropic structure was successfully reproduced in numerical models [25], up to now no clear explanation of this fact was given. We think that the present result relates quite clearly anisotropy in the interaction with anisotropy in the spatial arrangement.

We have provided rather direct evidence that the effective alignment interaction between starlings in a real flock is short-ranged. This result is interesting for two reasons. First, from a biological perspective, we believe this is the first time that 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 [22, 26–34]. 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. 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).

[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] A. Cavagna et al., Proc. Natl. Acad. Sci. 107, 11865 (2010).
[4] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, Proc. Natl. Acad. Sci. (2010).
[5] J. Gautrais et al., PLoS Comput. Biol 8, e1002678 (2012).
[6] A. Attanasi et al., ArXiv preprint , 1307.5631 (2013).
[7] D. J. T. Sumpter, J. Buhl, D. Biro, and I. D. Couzin, Theor. Biosci. 127, 177 (2008).
[8] A. Attanasi et al., ArXiv preprint , 1303.7097 (2013).
[9] M. Ballerini et al., Proc. Natl. Acad. Sci. 105, 1232 (2008).
[10] W. Bialek et al., Proc. Natl. Acad. Sci. 109, 4786 (2012).
[11] W. Bialek et al., Proc. Natl. Acad. Sci. 111, 7212 (2014).
[12] Y. Katz, K. Tunstrem, C. C. Ioannou, C. Huepe, and I. D. Couzin, Proc. Natl. Acad. Sci. 108, 18720 (2011).
[13] J. G. Puckett, D. H. Kelley, and N. T. Ouellette, Scientific reports 4, 4766 (2014).
[14] D. J. T. Sumpter, Collective Animal Behavior (Princeton University Press, New Jersey, 2010).
[15] 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).
[16] Note that the topological distance, $n$, is a discrete variable, hence the correlation function $C(n)$ is in fact a discrete set of values and it should be indicated as $C_n$. 
However, we prefer the $\bar{C}(n)$ notation, as it is reminiscent of the standard metric definition of correlation function, $C(r)$. For the same reason we inappropriately use Dirac’s notation $\delta(k_{ij} - n)$ for what is actually a Kronecker’s $\delta_{k_{ij}, n}$.

[17] A. Cavagna et al., Animal Behaviour 76, 217 (2008).
[18] A. Cavagna, I. Giardina, A. Orlandi, G. Parisi, and A. Proccaccini, Animal Behaviour 76, 237 (2008).
[19] A. Attanasi et al., ArXiv preprint , 1305.1495 (2013).
[20] E. T. Jaynes, Phys. Rev. 106, 620 (1957).
[21] M. Ballerini et al., Animal Behaviour 76, 201 (2008).
[22] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature 440, 1007 (2006).
[23] C. E. Shannon, Bell Sys. Tech. 27, 379 (1948).

We note that the same calculation 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 $r_s$ scales with the density, thus confirming the result of [9] that the interaction in starling flocks is based on topological distance.

[25] C. K. Hemelrijk and H. Hildenbrandt, Interface focus 2, 726 (2012).
[26] J. Shiels et al., J. Neurosci. 29, 8254 (2006).
[27] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, Proc. Natl. Acad. Sci. 103, 19033 (2006).
[28] A. Tang et al., J. Neurosci. 28, 505 (2008).
[29] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. 106, 67 (2009).
[30] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell 138, 774 (2009).
[31] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc. Natl. Acad. Sci. 107, 5405 (2010).
[32] G. Tkačik, ArXiv preprint . 1006:4291 (2010).
[33] G. Tkačik et al., J. Stat. Mech. 2013, P03011 (2013).
[34] G. Tkačik et al., PLoS Comput. Biol. 10, e1003408 (2014).

### SUPPLEMENTAL INFORMATION

#### I. MAXIMUM ENTROPY APPROACH: GENERAL PRINCIPLES

Consider a system whose microstate at any instant of time is described by a set of variables $\{x_1, x_2, ..., x_i, ..., x_N\} \equiv X$. In an experiment one draws $K$ independent microstates $X_1, X_2, ..., X_s, ..., X_K$ and measures the average values of some relevant quantities. Let us assume that several functions $f_1(X), f_2(X), ..., f_M(X)$ are measured; the experimental averages of these functions are denoted by $\langle f_1 \rangle_{\exp}, \langle f_2 \rangle_{\exp}, ..., \langle f_M \rangle_{\exp}$ respectively. The maximum entropy (ME) method consists in finding the most random probability distribution $P(X)$ that is consistent with the observed experimental data. The distribution must satisfy the following constraint equation:

$$\langle f_\mu \rangle_{\exp} = \langle f_\mu \rangle_{P}, \quad (8)$$

namely,

$$\frac{1}{K} \sum_{s=1}^{K} f_\mu(X_s) = \sum_X P(X)f_\mu(X), \quad (9)$$

for all $\mu = 1, 2, ..., M$, and where $\langle f_\mu \rangle_{P}$ denotes the expectation value computed using the probability distribution $P(X)$. Many distributions satisfy Eq. (9). The maximum entropy principle [1] aims to find the one which has as little structure as possible, so that one can derive the minimal consequences of the experimental observations on $\langle f_\mu \rangle_{\exp}$. The entropy is defined as $\mathcal{S} = -\sum_X P(X) \log P(X)$.

In order to get the desired probability distribution, we maximize $\mathcal{S}[P]$ under the constraints given by Eq. (9). Apart from the experimental constraints, there is an additional constraint, namely that the probability distribution should be normalized $\sum_X P(X) = 1$. This is equivalent to say that we add to our list of observables an extra function, the constant $f_0(X) = 1$. This constraint maximization problem can be solved with the Lagrange multiplier method [4] by finding the maximum of the generalized entropy function,

$$\mathcal{S}[P; \{\lambda_\nu\}] = \mathcal{S}[P] - \sum_{\mu=0}^{M} \lambda_\mu (\langle f_\mu \rangle_{P} - \langle f_\mu \rangle_{\exp}), \quad (11)$$

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

$$P(X) = \frac{1}{Z(\{\lambda_\nu\})} \exp \left[ - \sum_{\mu=1}^{M} \lambda_\mu f_\mu(X) \right], \quad (12)$$

where $Z(\{\lambda_\nu\})$ is obtained optimizing with respect $\lambda_0$ and it is given by,

$$Z(\{\lambda_\nu\}) = \exp(1 + \lambda_0) = \sum_X \exp \left[ - \sum_{\mu=1}^{M} \lambda_\mu f_\mu(X) \right]. \quad (13)$$

Then, the generalised entropy function (Eq. (11)) can be written as,

$$\mathcal{S}[P; \{\lambda_\nu\}] = \log Z(\{\lambda_\nu\}) + \sum_{\mu=1}^{M} \lambda_\mu \langle f_\mu \rangle_{\exp}. \quad (14)$$

Note that, maximizing with respect to $\lambda_\mu$, we recover the original constraint equation [9],

$$- \frac{\partial \log Z(\{\lambda_\nu\})}{\partial \lambda_\mu} = \sum_X P(X)f_\mu(X) = \langle f_\mu \rangle_{\exp}. \quad (15)$$

In other words, the maximum entropy method produces a Boltzmann distribution $P(X) = \exp(-\beta H(X))/Z$ for
a system with “Hamiltonian” \( H(X) = \sum_{\mu=1}^{M} \lambda_{\mu} f_{\mu}(X) \) in thermal equilibrium at temperature \( k_B T = 1/\beta = 1 \) and, accordingly, \( Z(\{\lambda_{\nu}\}) \) is exactly the partition function of the system.

Equivalently, one can recover \( \{\lambda_{\nu}\} \) also by maximizing the log-likelihood function \( \log Z(\{\lambda_{\nu}\}) \) of the probability distribution \( P(X) \). For a given sample \( (X_1, X_2, ..., X_K) \), the log-likelihood can be written as,

\[
\log Z(\{\lambda_{\nu}\}) = \left( \sum_{s=1}^{K} \log P(X_s) \right) = -n \log Z(\{\lambda_{\nu}\}) - \sum_{\mu=1}^{M} \lambda_{\mu} \langle f_{\mu} \rangle_{\text{exp}}. \tag{16}
\]

Note that the log-likelihood function is minus the entropy (Eq. 14) and maximizing the log-likelihood function one also can recover constraint equations (Eq. 15).

II. MAXIMUM ENTROPY FOR FLOCKS

We explain here how to apply the ME method to experimental data about bird flocks. We will first consider the case of interactions that only depend on the order of neighborhood (no angular dependence) and secondly we will generalize the method to the anisotropic case.

A. Isotropic interaction

The experimental observable we consider in flocks is the correlation function \( C(n) \). In principle, this correlation functions can be computed for each value of the topological distance \( n \) but, as will discuss later, it is safe to consider only the observables \( C(n) \) up to \( n = n_{\text{max}} \ll N \). 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 \( C(n) \) contributions from the distances \( n, n+1, ..., n+\Delta n - 1 \). Hence, we have,

\[
\hat{C}(n) = \frac{\sum_{i,j} \delta(k_{ij} - n) |s_i \cdot s_j|}{\sum_{i,j} \delta(k_{ij} - n)},
\]

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

\[
\delta(k-n) = \begin{cases} 1 & \text{if } n \leq k < n + \Delta n, \\ 0 & \text{otherwise}. \end{cases} \tag{17}
\]

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

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

\[
H = \sum_n \lambda_n \hat{C}(n), \tag{18}
\]

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

In fact, by defining \( \hat{J}(k_{ij}) \) such that

\[
\hat{J}(k_{ij}) = \sum_n J(n) \delta(k_{ij} - n), \tag{19}
\]

(note that with \( \Delta n = 1 \), \( \hat{J}(k_{ij}) \equiv J(k_{ij}) \)), we can easily verify that Eq. 18 reads as classical Heisenberg model

\[
H(\{s_i\}) = -\sum_{i,j} \hat{J}(k_{ij}) s_i \cdot s_j \equiv -\sum_{i,j} J_{ij} s_i \cdot s_j. \tag{20}
\]

The \( \{s_i\} \) are 3-dimensional spins and \( \hat{J}(k_{ij}) \) are the intensities of the pairwise interactions. It is convenient to introduce the symmetrized interactions matrix \( J_{ij} \)

\[
J_{ij} = \frac{1}{2} [\hat{J}(k_{ij}) + \hat{J}(k_{ji})]
\]

\[
= \frac{1}{2} \sum_n J(n) [\delta(k_{ij} - n) + \delta(k_{ji} - n)], \tag{21}
\]

since it will be useful to derive analytically expression of partition function \( Z \) and its derivatives in a simpler way. In this way, \( J(n) \) will measure how a generic bird interacts with its \( n^{th} \) neighbour.

B. Longitudinal vs transverse interaction

As we discuss in the main text we divide the space around a bird into two sectors, the longitudinal one (L) and the transversal one (T) to study the anisotropy in the alignment interaction. The observables suitable for this purpose are the following,

\[
\hat{C}^L(n) = \frac{\sum_{i,j} \delta(k_{ij} - n) \Theta(\cos(\theta_{ij})) - \frac{1}{2}}{\sum_{i,j} \delta(k_{ij} - n) \Theta(\cos(\theta_{ij})) - \frac{1}{2}},
\]

\[
\hat{C}^T(n) = \frac{\sum_{i,j} \delta(k_{ij} - n) \Theta(1/2 - |\cos(\theta_{ij})|)}{\sum_{i,j} \delta(k_{ij} - n) \Theta(1/2 - |\cos(\theta_{ij})|)},
\]

where \( \theta_{ij} \) is the angle formed by \( \vec{r}_{ij} = (\vec{r}_j - \vec{r}_i) \) and the flock’s direction of motion \( \vec{V} \),

\[
\cos \theta_{ij} = \frac{\vec{r}_{ij}}{|\vec{r}_{ij}|} \frac{\vec{V}}{|\vec{V}|} \quad , \quad \vec{V} = \frac{1}{N} \sum_i \vec{s}_i.
\]
\(\Theta(x)\) is the Heaviside step function and the factor 1/2 divides the space evenly between the two sectors. The \(\delta\)-function bears the same meaning as in the section above (Eq. (17)).

Following the same method as above, we introduce the Lagrange multipliers \(\lambda_n^{L,T}\), and we write down the Hamiltonian

\[
H = \sum_n \lambda_n^{L} \hat{C}_n^{L}(n) + \lambda_n^{T} \hat{C}_n^{T}(n)
\]

We define the fraction of neighbors that lie in longitudinal and transversal sector for each \(n\),

\[
p^L(n) = \frac{1}{N_\Delta n} \sum_{i,j} \delta(k_{ij} - n) \Theta(|\cos(\theta_{ij})| - 1/2),
\]

\[
p^T(n) = \frac{1}{N_\Delta n} \sum_{i,j} \delta(k_{ij} - n) \Theta(1/2 - |\cos(\theta_{ij})|),
\]

(22)

These quantities of course satisfy the relation \(p^L(n) + p^T(n) = 1\). The longitudinal and transversal interaction are obtained by substituting \(J^{L,T}(n) = -\lambda_n^{L,T}/(p^L(n)N_\Delta n)\). Then, as above, we introduce the pairwise interactions \(\hat{J}_n^{L,T}(k_{ij})\)

\[
\hat{J}_n^{L,T}(k_{ij}) = \sum_n \lambda_n^{L,T} \delta(k_{ij} - n) \Theta_{ij}^{L,T},
\]

(23)

where \(\Theta_{ij}^{L,T} = \Theta(\pm |\cos(\theta_{ij})| \mp 1)\) and it selects for a generic bird \(i\) and its neighbors \(j\) only one interaction between \(J_T(k_{ij})\) and \(J_L(k_{ij})\). With these substitutions the Hamiltonian is the same as in the isotropic case (Eq. (20)),

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

where, thanks to \(\Theta_{ij}^{L,T} = \Theta_{ji}^{L,T}\), we can symmetrize again the \(\hat{J}_n^{L,T}(k_{ij})\) into \(J_{ij}\)

\[
J_{ij} = \frac{1}{2}(\hat{J}_n^{L}(k_{ij}) + \hat{J}_n^{L}(k_{ij}^*)) + \frac{1}{2}(\hat{J}_n^{T}(k_{ij}) + \hat{J}_n^{T}(k_{ij}^*))
\]

\[
= \frac{1}{2} \sum_n \lambda_n^{L,T} [\delta(k_{ij} - n) + \delta(k_{ij} - n)] + \frac{1}{2} \sum_n \lambda_n^{L,T} [\delta(k_{ij} - n) + \delta(k_{ij} - n)].
\]

(24)

Note that, despite the constrains:

\[
\hat{C}(n) = p^L(n)\hat{C}_n^{L}(n) + p^T(n)\hat{C}_n^{T}(n),
\]

the link between \(J(n)\) and \(J^{L,T}(n)\) is not trivial. In this case we match at the same time \(\hat{C}_n^{L}(n)\) and \(\hat{C}_n^{T}(n)\), in the isotropic case \(\hat{C}(n)\) only. But \(\hat{C}_n^{L}(n)\) and \(\hat{C}_n^{T}(n)\) can be very different from each other while maintaining \(\hat{C}(n)\) constant (Eq. (25)). That means that we can have many combinations of \(J^{L,T}(n)\) compared to the one for \(J(n)\).

Special cases occur only when the correlation functions of the two sectors are the same \(\hat{C}_n^{L}(n) = \hat{C}_n^{T}(n)\) (in this case \(J_L(n) = J_T(n) = J(n)\)) or when there are no neighbors in one of sectors (if \(p^L(n) = 1\) then \(J_T(n) = J(n)\) and \(J_L(n)\) is indeterminate, and vice versa), as it should be.

In conclusion, \(J_L(n)\) and \(J_T(n)\) measure the interaction between a generic bird and its \(n^{th}\) neighbor according to the sector occupied by such neighbor.

### C. Partition function in spin-wave approximation

In order to calculate the log-likelihood we need to compute the partition function \(Z\) (Eq. (10)). In general, the exact analytical calculation of the partition functions is very hard. In the case of flock, however, this can be done thanks to the spin-wave approximation. The idea is that, because flocks are very ordered, with magnetization (i.e. polarization) close to 1, we can expand in the small fluctuation around the mean direction of motion.

The partition function can be written as,

\[
Z = \int D\vec{\pi} \exp \left[ \sum_{i,j} J_{ij} \vec{s}_i \cdot \vec{s}_j \right]
\]

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

\[
\vec{\pi}_i \cdot \hat{n} = 0 \quad \text{and} \quad \frac{1}{N} \sum_i s_i^T = \Phi \quad \text{and} \quad \sum_i \vec{\pi}_i = 0.
\]

The \(\vec{\pi}_i\) are the small fluctuations that the spin-wave approximation uses to expand the partition function. The partition function can be rewritten as,

\[
Z = \int Ds^L D\vec{\pi} \prod_i \delta \left( \sqrt{(s_i^L)^2 + |\vec{\pi}_i|^2} - 1 \right) \times
\]

\[
\times \delta \left( \sum_i \vec{\pi}_i \right) \exp \left[ \sum_{i,j} J_{ij} (s_i^L s_j^L + \vec{\pi}_i \cdot \vec{\pi}_j) \right]
\]

(24)

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 flock, \(\Phi \approx 1\) and \(|\vec{\pi}_i| \ll 1\). Then, at second order, \(s_i^L = 1 - |\vec{\pi}_i|^2/2\). Performing the integral over \(s_i^L\), the partition function becomes,

\[
Z = \int D\vec{\pi} \prod_i \frac{1}{\sqrt{1 - |\vec{\pi}_i|^2}} \delta \left( \sum_i \vec{\pi}_i \right) \times
\]

\[
\times \exp \left[ -\sum_{i,j} A_{ij} \vec{\pi}_i \cdot \vec{\pi}_j + \sum_i J_{ij} \right],
\]
where

\[ A_{ij} = \delta_{ij} \left( \sum_k J_{ik} \right) - (1 - \delta_{ij})J_{ij}. \]  

(26)

For strongly order flock, the product \( \prod \sqrt{1 - |\vec{w}_i|^2} \) can be neglected (we have explicitly checked that the corrections due to this term are indeed negligible). Therefore we can write

\[
Z = \int D\vec{\pi} \exp \left[ -\sum_i A_{ij} \vec{\pi}_i \cdot \vec{\pi}_j + \sum_{ij} J_{ij} \right].
\]

Because \( J_{ij} = J_{ji} \) (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, in particular, the smallest eigenvalue is \( \gamma_1 = 0 \), and all other eigenvalues are positive. Let \( \gamma_j \) be the eigenvalue of the eigenvector \( \vec{w}_j \).

The eigenvector \( \vec{w}_k \) satisfy the usual relation,

\[
\sum_j A_{ij} w_j^k = \gamma_k w_i^k.
\]

(27)

It can be easily seen that the eigenvector \( \vec{w}_1 \) corresponding to \( \gamma_1 \) is constant and it is given by \( (1/\sqrt{N}, 1/\sqrt{N}, ..., 1/\sqrt{N}) \). We can rewrite the integral in orthonormal basis defined by \( \vec{w}_1, \vec{w}_2, ..., \vec{w}_N \):

\[
Z = \int D\vec{\pi} \exp \left[ -\sum_{i,j} J_{ij} \right],
\]

where \( \vec{\pi}_k = \sum_i w_i^k \vec{\pi}_i \). From this form it is clear that the role of the \( \delta \)-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,

\[
\log Z = -\sum_{k>1} \log a_k + \sum_{ij} J_{ij},
\]

(28)

where the irrelevant constant terms have been neglected.

**D. Maximizing the log-likelihood**

Using the analytical expression for partition function Eq. (27), we can write the expressions of the log-likelihood (Eq. (10)) for the case of full interaction,

\[
\log L = \sum_{k>1} \log a_k + N\Delta n \sum_n J(n) \langle \hat{C}(n) \rangle_{\text{exp}} = \sum_{k>1} \log a_k - N\Delta n \sum_n J(n) (1 - \langle \hat{C}(n) \rangle_{\text{exp}}).
\]

(29)

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

\[
\log L = \sum_{k>1} \log a_k + N\Delta n \sum_n J(n) (1 - \langle C_T(n) \rangle_{\text{exp}}) + \sum_{k>1} \log a_k - N\Delta n \sum_n J(n) (1 - \langle C_T(n) \rangle_{\text{exp}}).
\]

(29)

The condition for maximizing the log-likelihood is \( \partial \log L / \partial J(n) = 0 \). Let us consider, for the moment, the isotropic case. The derivatives 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 \( \gamma_k \) are very complicated functions of the \( \{J(n)\} \). We will calculate \( \partial \gamma_k / \partial J(n) \) by using perturbation theory. Suppose that we perturb \( J(n) \) by some small amount, \( J(n) \rightarrow J(n) + \epsilon \), where, \( \epsilon \) is infinitesimal. The perturbation makes \( A_{ij} \) change into,

\[
\tilde{A}_{ij}(\epsilon) = A_{ij} + \epsilon \gamma_{ij}(n),
\]

where we introduced a symmetric matrix

\[
\gamma_{ij}(n) = \frac{\partial A_{ij}}{\partial J(n)} = \sum_{lm} \frac{\partial A_{ij}}{\partial J_{lm}} \frac{\partial J_{lm}}{\partial J(n)}.
\]

(30)

Due to this small perturbation the eigenvalue \( \gamma_k \) and its eigenvector \( \vec{w}_k \) change by small amount,

\[
\tilde{\gamma}_k = \gamma_k + \epsilon \xi_k + O(\epsilon^2),
\]

\[
\tilde{\vec{w}}_k = \vec{w}_k + \epsilon g_k + O(\epsilon^2).
\]

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

\[
\sum_j \tilde{A}_{ij}(\epsilon) \tilde{\vec{w}}_j(\epsilon) = \tilde{\gamma}_k(\epsilon) \tilde{\vec{w}}_k(\epsilon).
\]

Through some algebra it is quite straightforward to show that,

\[
\tilde{\gamma}_k(\epsilon) = \gamma_k + \epsilon \sum_{ij} \gamma_{ij}(n) w^k_i w^k_j + O(\epsilon^2).
\]

Therefore, the derivative of the eigenvalue \( \gamma_k \) can be written as

\[
\frac{\partial \gamma_k}{\partial J(n)} = \lim_{\epsilon \to 0} \frac{\tilde{\gamma}_k(\epsilon) - \gamma_k}{\epsilon} = \sum_{ij} \gamma_{ij}(n) w^k_i w^k_j.
\]

(30)

To obtain the form of matrix \( \gamma_{ij}(n) \) we use first Eq. (26) and Eq. (21) from which

\[
\frac{\partial A_{ij}}{\partial J_{lm}} = \delta_{il} (\delta_{jm} - (1 - \delta_{ij})\delta_{jm})
\]

\[
\frac{\partial J_{lm}}{\partial J(n)} = \frac{1}{2} [\delta(k_{lm} - n) + \delta(k_{ml} - n)]
\]

(30)
then from Eq. (30)
\[
\gamma_{ij}(n) = \frac{1}{2} \delta_{ij} \left[ \sum_m (\delta(k_{im} - n) + \delta(k_{mi} - n)) \right] + \\
- \frac{1}{2} (1 - \delta_{ij}) (\delta(k_{ij} - n) + \delta(k_{ji} - n)).
\]

Similarly, using Eq. (24), we obtain \( \gamma_{ij}^{L,T}(n) \) for the anisotropic case,
\[
\gamma_{ij}^{L,T}(n) = \frac{1}{2} \delta_{ij} \left[ \sum_m (\delta(k_{im} - n) + \delta(k_{mi} - n)) \Theta_{im}^{L,T} \right] + \\
- \frac{1}{2} (1 - \delta_{ij}) (\delta(k_{ij} - n) + \delta(k_{ji} - n)) \Theta_{ij}^{L,T}.
\] (32)

Now, with Eq. (31), it becomes easy to calculate the derivatives of the log-likelihood Eq. (28) respect each its variables \( J(n) \) and imposing its maximization we obtain
\[
1 - \langle \hat{C}(n) \rangle_{\exp} = \frac{1}{N \Delta n} \sum_{k>1} \frac{1}{a_k} \frac{\partial a_k}{\partial J(n)} = \\
= \text{Tr}[A^{-1} \gamma(n)]
\] (33)

Similarly, in the anisotropic case the maximization of Eq. (29) gives,
\[
1 - \langle \hat{C}^{L,T}(n) \rangle_{\exp} = \frac{1}{\theta^{L,T}(n) N \Delta n} \sum_{k>1} \frac{1}{a_k} \frac{\partial a_k}{\partial J^{L,T}(n)} = \\
= \frac{\text{Tr}[A^{-1} \gamma^{L,T}(n)]}{\theta^{L,T}(n) N \Delta n}
\] (34)

E. Stability of the 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. (28), Eq. (29)). 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 some numerical method.

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 iteration needed for the optimization is much larger than in the second case, as we provide less information. Furthermore, as the dimension of the log-likelihood function gets larger, the number of iterations in the first case increases very rapidly. During each iteration step the most time-consuming part is the diagonalization of the matrix \( A_{ij} \). Because the number of iterations is significantly smaller with method (ii) than with (i), the whole computation gets much more efficient. 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 multimizer nnmsimplex2, belonging to the gnu scientific library [6]. This optimization algorithm is based on Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [7]. 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 [6].

In Fig. 4 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 our solution is very stable and it is also reached very quickly.

III. COMPARISON WITH THE STEP-INTERACTION CASE

The first maximum entropy calculation on bird flocks was different from the present one; in particular, concerning the range and the nature of the interaction it was not really a maximum entropy calculation, but a maximum likelihood one. The experimental input used in the integrated correlation function over a (topological) range \( n_c \),
\[
\hat{C}_{int}(n_c) = \frac{1}{N n_c} \sum_{j \in i} \bar{s}_i \cdot \bar{s}_j
\] (35)
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,

\[
\hat{C}_{\text{int}}(n_c) = \frac{\sum_{i,j=1}^{N} \hat{s}_i \cdot \hat{s}_j \Theta(k_{ij} - n)}{\sum_{i,j=1}^{N} \Theta(k_{ij} - n)},
\]

where \( \Theta(x) \) is the Heaviside step function. By using the formalism that we developed above, it is easy to see that this construction assumes that the interaction strength is constant up to neighbour \( n_c \) and zero beyond that, namely,

\[
J(n) = J_0 \Theta(n_c - n).
\]

Hence, \( J_0 \) was the (average) strength of the interaction and it was the Lagrange multiplier associated to \( C_{\text{int}} \), so that the entropy was maximized w.r.t. \( J_0 \). On the other hand, \( n_c \) was the width of the step-like interaction, but this parameter was not the Lagrange multiplier of any given observable, so it remained in the likelihood even after maximization w.r.t. \( J_0 \). For this reason, \( n_c \) was determined through a maximum likelihood principle.

This means that the calculation of \( n_c \) in fact assumed some parameter-dependent form of the model (namely the step interaction form \( (36) \)), hence it was not completely unbiased.

How does the calculation of \( n_c \) compare to the one we developed above? First of all, we see from Fig. 5 that the old step-interaction is always quite close to the new exponential interaction, so there is a nice consistency between the two cases. For a more quantitative comparison, let us call \( J_0^{\text{step}} \) and \( n_c^{\text{step}} \) the strength and the range of the interaction of the step model of \( [10] \), and \( J_0^{\exp} \) and \( n_c^{\exp} \) the parameters of the exponential fit of the \( J(n) \) that we calculated in the present work (Eq. (7)).

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 \( n_c^{\exp} J_0^{\exp} = n_c^{\text{step}} J_0^{\text{step}} \);

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

FIG. 6. (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 \( [10] \).
These two conditions give,

\[ n_c^{\exp} = n_c^{\text{step}} / 2, \]
\[ J_0^{\exp} = 2 J_0^{\text{step}}. \]  

(38)

The data in Fig. 6 suggest that these two relations are indeed satisfied. Notice that the fact that the step-like parameter \( n_c^{\text{step}} \) is twice as large as the exponential decay range can (at least partially) explain the discrepancy between \( n_c^{\exp} \) and the previous estimate of the interaction range given in [3].

IV. CHOOSING THE MAXIMUM NUMBER OF PARAMETERS

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 our work we use the correlation function \( C(n) \) as observable, which is in fact a list of \( n_{\text{max}} \) numbers, with \( n = 1, \ldots, n_{\text{max}} \). Given one bird, the largest value of the topological distance is \( N \) (to be precise it is \( N - 1 \)), so that in principle we could use up to \( n_{\text{max}} = N \) observables. As a consequence, we should infer \( N \) Lagrange multipliers. In practice, this is not necessary and we stop at a maximum value \( n_{\text{max}} = 23 \), because we have seen in the main text that nothing changes in the form of \( J(n) \) by adding extra parameters. Here, we want to discuss the issue of the maximum number of parameters a bit more carefully.

In Fig. (7a), we have plotted the entropy vs \( n_{\text{max}} \) for a flock (one snapshot). We see that the entropy decreases with \( n_{\text{max}} \): when we increase \( n_{\text{max}} \), we add experimental information and as a consequence the entropy never ceases to decay. Basically, this is telling us that the fit gradually improves by adding more parameters. However, this decrease of the entropy is uneven: it initially decreases fast, but after a certain \( n_{\text{max}} \approx 10 \), it starts decaying slowly, and quite linearly.

To understand this change of behaviour of the entropy we need a Bayesian analysis. We must recall that the goal of our work is to find the minimal set of parameters able to reproduce the experimental data. The probability \( P(n_{\text{max}} | D) \) of a model with \( n_{\text{max}} \) parameters, given a certain dataset \( D \), is the product of two terms: the first term is the maximized likelihood; the logarithm of its inverse it is the entropy (Eq. (16)). The second term, which is called Occam factor, \( V(n_{\text{max}}) \), is equal to the ratio between the posterior accessible volume in the space of parameters and the prior accessible volume \( V \). Typically, the Occam factor decays exponentially with the number of parameters, \( V(n_{\text{max}}) \propto e^{-\alpha n_{\text{max}}} \).

Hence, in general when we increase the number of parameter \( n_{\text{max}} \) 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 \( P(n_{\text{max}} | D) \) reaches a maximum for a finite value of parameters, \( n_{\text{max}} = n_{\text{opt}}^{\text{max}} \), Fig. 7(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 \( n_{\text{opt}}^{\text{max}} \). First of all, the contribution of the Occam factor to \( P(n_{\text{max}} | D) \) depends only logarithmically on the prior probability: major changes in the prior leads to a small change in Occam factor. Moreover, when we increase \( n_{\text{max}} \) after we reach the optimal value, the slope of entropy changes suddenly: in fact, we move from the regime \( n_{\text{max}} < n_{\text{opt}}^{\text{max}} \), where adding each new observable implies a considerable increase of information, to the regime \( n_{\text{max}} > n_{\text{opt}}^{\text{max}} \), where instead adding a new observables only marginally increase the total information. \( n_{\text{opt}}^{\text{max}} \) is determined by the condition \( \partial S / \partial n_{\text{max}} = \partial \log V(n_{\text{max}}) / \partial n_{\text{max}}, \) which means that the solution is the crossing point between red and blue line in Fig. 7(c). Varying the prior, the blue line moves up and down and this moves \( n_{\text{opt}}^{\text{max}} \) by an amount \( \Delta n_{\text{max}} \).

As we can see from the figure, the faster the change of slope of entropy, the smaller the range \( \Delta n_{\text{max}} \). Then, typically, big changes in the prior leads to little change in \( n_{\text{opt}}^{\text{max}} \).

V. DATA SET

The 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 [8, 10–15]. 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 was taken in the period 2005-2008, with cameras shooting at 10 frames-per-second (fps). The top 18 events in Table I (21-06 to 77-07) belong to the first set. The data for the second set was collected in the period between 2010-2012, and for these events were used cameras shooting at 170fps [14, 15]. The last 4 events in Table I belong to the second set. All the events correspond to strongly ordered flocks, with polarization between $\Phi = 0.844$ and $\Phi = 0.995$, hence justifying the spin wave expansion.

[1] E. T. Jaynes, Phys. Rev. 106, 620 (1957).
[2] C. E. Shannon, Bell Sys. Tech. 27, 379 (1948).
[3] T. M. Cover and J. Thomas, Elements of Information Theory (Wiley, New York, 1991).
[4] C. M. Bender and S. A. Orszag, Advanced Mathematical Methods for Scientists and Engineers (McGraw-Hill, New York, 1978).
[5] W. Bialek et al., Proc. Natl. Acad. Sci. 109, 4786 (2012).
[6] B. Gough, GNU Scientific Library Reference Manual - Third Edition, 3rd ed. (Network Theory Ltd., 2009).
[7] R. Fletcher, Practical Methods of Optimization, 2nd ed. (Wiley, 1987).
[8] M. Ballerini et al., Proc. Natl. Acad. Sci. 105, 1232 (2008).
[9] D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms (Cambridge University Press, 2003).
[10] A. Cavagna et al., Animal Behaviour 76, 217 (2008).
[11] A. Cavagna, I. Giardina, A. Orlandi, G. Parisi, and A. Procaccini, Animal Behaviour 76, 237 (2008).
[12] M. Ballerini et al., Animal Behaviour 76, 201 (2008).
[13] A. Cavagna et al., Proc. Natl. Acad. Sci. 107, 11865 (2010).
[14] A. Attanasi et al., ArXiv preprint , 1303.7097 (2013).
[15] A. Attanasi et al., ArXiv preprint , 1305.1495 (2013).

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

Table I. Flocks Data: Each line represents a different flocking event. $N$ is the number of individuals in the flock, $\Phi$ the average polarization, $L$ the size of the flock (maximum distance between two birds), $n_{exp}$ the exponential decay range computed in this work and $n_{step}$ the range of interaction of step model of [5].