

<https://doi.org/10.1038/s42005-022-00864-9>

OPEN

Inferring the size of a collective of self-propelled Vicsek particles from the random motion of a single unit

Pietro De Lellis ¹✉ & Maurizio Porfiri ²✉

Inferring the size of a collective from the motion of a few accessible units is a fundamental problem in network science and interdisciplinary physics. Here, we recognize stochasticity as the commodity traded in the units' interactions. Drawing inspiration from the work of Einstein-Perrin-Smoluchowski on the discontinuous structure of matter, we use the random motion of one unit to identify the footprint of every other unit. Just as the Avogadro's number can be determined from the Brownian motion of a suspended particle in a liquid, the size of the collective can be inferred from the random motion of any unit. For self-propelled Vicsek particles, we demonstrate an inverse proportionality between the diffusion coefficient of the heading of any particle and the size of the collective. We provide a rigorous method to infer the size of a collective from measurements of a few units, strengthening the link between physics and collective behavior.

¹Department of Electrical Engineering and Information Technology, University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy. ²Department of Mechanical and Aerospace Engineering, Department of Biomedical Engineering, and Center for Urban Science and Progress, New York University Tandon School of Engineering, Brooklyn, NY 11201, USA. ✉email: pietro.delellis@unina.it; mporfiri@nyu.edu

Collective dynamics are ubiquitous in nature^{1,2}. From neural circuits to animal groups, there are countless instances where the interactions among large numbers of elementary units bestow surprisingly complex patterns of tantalizing beauty for the collective. Deciphering the interactions within a collective from observations of its units is the object of several successful methodological endeavors^{3,4}. Common to the vast majority of the existing approaches is the knowledge of the size of the collective, a hypothesis that is often defeated in experiments. For example, understanding bird migration requires tagging individuals to measure their motion⁵, but there is no guarantee that these tagged individuals comprise the totality of the flock.

Only a handful of recent studies have addressed the fundamental problem of estimating the size of a collective from measurements of some of its units^{6–9}. The approaches proposed by Haenhe et al.⁶, Porfiri⁷, and Tang et al.⁸ are based on the computation of the rank of ancillary matrices that encode the transient response of some of the units, in the absence of forcing. In contrast, Tyloo and Delabay⁹ tackle the inference problem from steady-state response to harmonic forcing. Although they may tolerate some level of added noise and measurement uncertainty, all of these methods are designed on the premise of completely deterministic dynamics.

Here, we propose a different paradigm to solve the problem of estimating the size of a collective, building upon seminal ideas of Einstein¹⁰, Smoluchowski¹¹, and Perrin¹² that led to the determination of the Avogadro's number from kinetic theory. Using the words of Feynman regarding the experiments of Perrin¹³, “one of the earliest determinations of the number of atoms was by the determination of how far a dirty little particle would move if we watched it patiently under a microscope for a certain length of time.” We demonstrate that the mean square dynamics of a single unit is sufficient to infer the size of a collective.

We focus on a system of self-propelled Vicsek particles¹⁴ as a universal model for collective dynamics, for which we show that the time rate of growth of the mean square heading of any particle is sufficient to predict the number of particles in the system. The concept is best explained by drawing an analogy between Vicsek models and the classical kinetic theory of gases¹⁵, where the mean square displacement of any particle is given by $\lambda \propto Dt$, with t being time and D the diffusion coefficient. The latter can be estimated as $D \propto \frac{k_B T}{m N d^2}$, where k_B is Boltzmann's constant, T the absolute temperature, V the volume of the gas, d the particle diameter, m the particle mass, and N the number of particles. Hence, one can estimate the number of particles in the gas from the rate of growth of the mean square displacement, upon some knowledge of the thermodynamic state of the gas and the properties of its particles.

Provided that the network of interactions underlying the collective dynamics of the Vicsek particles guarantees some average connectivity in time, the motion of one unit will contain a significant footprint of the size of the entire system, similar to existing approaches for the inference of the size of a collective^{6–9}. We numerically and mathematically demonstrate that the inference of the size of the system from the motion of a single particle only requires knowledge about the variance of the added noise, which, indeed, has been proposed as a thermodynamic analog of the temperature for collective dynamics^{16–19}. Different from the experiments of Perrin, our focal unit is not a macroscopic particle in a bath of microscopic particles, rather, it is one of the very particles comprising the collective whose trajectory is accessible by the experiment.

Results

Numerical results. We consider a system of N self-propelled particles in a square domain of size $L \times L$ with periodic boundary

conditions. At time $k \in \mathbb{Z}_{\geq 0}$, the planar position of particle i is $r_k^i \in \mathbb{C}$ and its heading is $\theta_k^i \in [-\pi, \pi]$. Particles negotiate their heading with other particles within a unitary interaction radius under the effect of additive noise, moving at a constant speed,

$$\begin{aligned}\theta_{k+1}^i &= \arg\left(\sum_{j \in \mathcal{N}_k^i} \exp(i\theta_k^j)\right) + \psi_k^i, \\ r_{k+1}^i &= r_k^i + s \exp(i\theta_k^i).\end{aligned}\quad (1)$$

Here i is the imaginary unit; $\mathcal{N}_k^i = \{j = 1, \dots, N : \|r_k^j - r_k^i\| \leq 1\}$ is the set of neighbors of particle i at time k ; $\psi_k^1, \dots, \psi_k^N$ are independent uniformly distributed random variables in $[-\eta/2, \eta/2]$ with $\eta \in \mathbb{R}_{>0}$ being a control parameter that modulates the added noise so that the noise variance is $\sigma^2 = \eta^2/12$; and $s \in \mathbb{R}_{>0}$ is the common speed of all the particles, which corresponds to the number of interaction radii that are traveled in one time-step, so that neighbors are almost fixed for $s \ll 1$ and they randomly mix for $s \gg 1$. Parameters N and L control the density of the system, η acts as a temperature, and s regulates temporal correlations among particles.

We report simulation results for two values of s (low and high), and several values of L and N , spanning the simulation range considered by Vicsek et al.¹⁴. For each parameter combination, we perform 1000 repetitions with the same initial configuration, randomly selected in a square of unit length at the center of the domain. The initial heading of each particle is randomly selected in $[-\pi, \pi]$, and η is equal to 0.1, which corresponds to the “small noise” case originally considered by Vicsek¹⁴ to demonstrate ordered motion in the collective (with an average velocity, sometimes called polarization, close to one). In each repetition, we run model (1) for 150 time steps and track the heading of a single (randomly selected) focal particle, see Fig. 1. For each parameter combination and at every time-step k , we compute the sample variance Y_k of the heading across the 1000 repetitions (for moderate noise, the sample variance is, in fact, equivalent to the sample circular variance). As one should expect from kinetic theory, Y_k grows linearly with time (that is, $Y_k = Dk + Y_0$, with Y_0 being the initial variance) and the diffusion coefficient (estimated using linear regression on Y_1, \dots, Y_{150}) is inversely proportional to the number of particles N , see Fig. 1.

Irrespective of the particle speed, there is a wide region of the parameter space (L^2, N) for which the diffusion coefficient per unit noise variance (D/σ^2 ; calculated by taking the mean of the diffusion coefficients estimated from all the particles) is equal to the inverse of the number of particles, see Fig. 2a, b. As a result, it is possible to exactly infer the size of the collective from knowledge of the heading of a focal particle and of the noise variance that captures the system temperature (for details on the precision of the estimate, see Supplementary Note 1). More specifically, for the low-speed case, the accuracy of the inference is robust with respect to changes in the domain size and the number of particles, with variations of at most 20% registered in the whole parameter space. For the high-speed case, we find a more marked dependence on the accuracy of the inference on L and N . Although accurate inferences (within 20%) are obtained for $L^2 \leq 40$, further increasing the domain size challenges the inference. Such a detrimental effect is more evident for small values of N . The possibility of accurately performing the inference across widely different particle speeds is surprising, whereby the Vicsek model spans from the diluted XY-model ($s \ll 1$) to the mean-field behavior of a ferromagnet ($s \gg 1$) as a function of the speed¹⁴. We offer some mathematical backing to this observation through the analytical treatment of these two limit cases, for which we establish closed-form expressions for the diffusion coefficient.

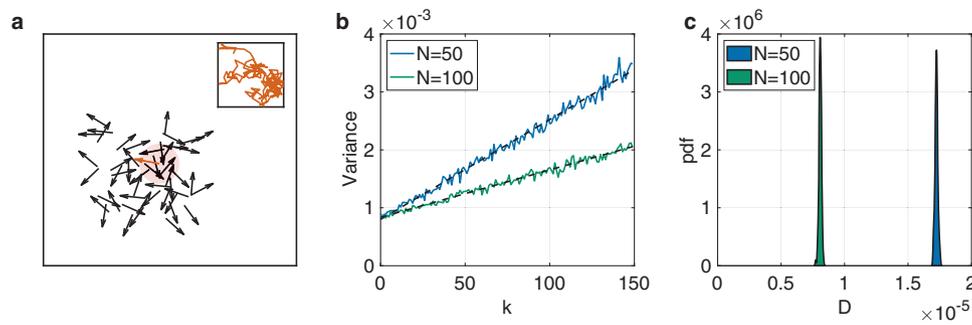


Fig. 1 Illustration of the proposed approach to infer the size of a collective of self-propelled Vicsek particles from the motion of a single unit. We perform 1000 repetitions where we record the heading of one particle in the collective for different values of the side length L of the square domain, speed s , number of particles N , and noise η . **a** Schematic of a numerical experiment for $N = 20$, where we display a time snapshot of the system (the focal particle is in red and its interaction circle is shaded). The inset depicts a sample trajectory of the focal particle evolving from the snapshot for 150 time steps. **b** Variance Y_k of the heading of the focal particle as a function of time k for two system sizes when $\eta = 0.1$, $s = 3$, and $L = 4$, with the dashed black lines marking linear fitting. Doubling the size halves the diffusion coefficient ($N = 50$: $D = 1.67 \times 10^{-5}$, and $N = 100$: $D = 8.41 \times 10^{-6}$). **c** Distribution of the diffusion coefficient as it would be estimated from observations of different focal particles in the collective.

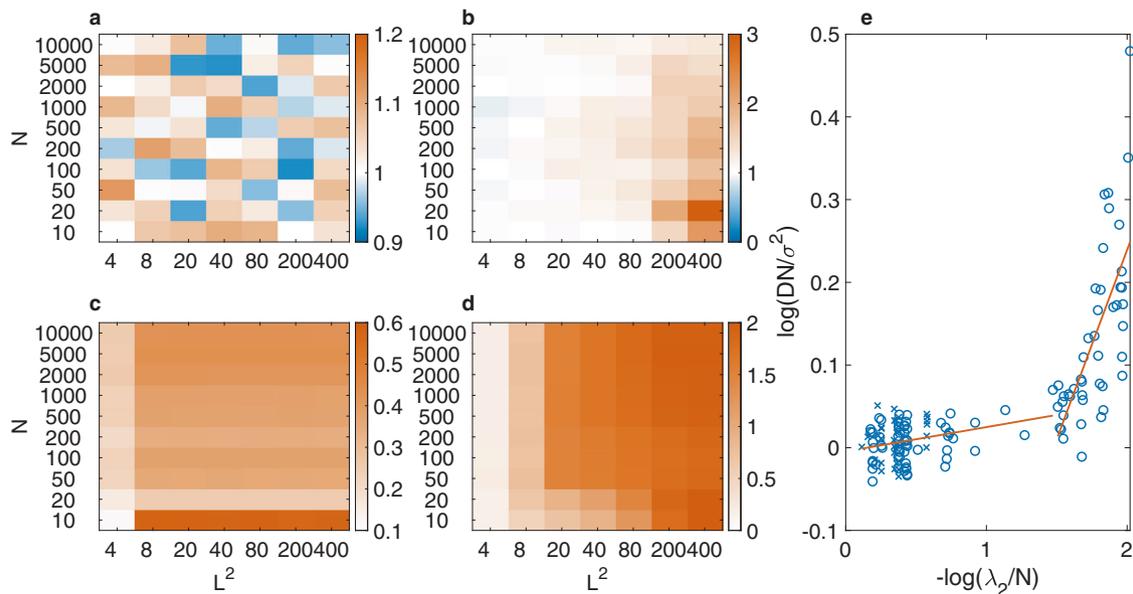


Fig. 2 Computational validation of the proposed inference approach. **a, b** Color maps of DN/σ^2 , where D is the diffusion coefficient and σ^2 is the noise variance, as a function of the side length L of the square domain and the system size N for $\eta = 0.1$ (**a, b** correspond to speed $s = 0.3$ and $s = 3$, respectively, see Supplementary Note 5 for details on the speed selection). Perfect inference of the size of the collective ($DN/\sigma^2 = 1$) is indicated in white, the darker the red the worse is the underestimation ($DN/\sigma^2 > 1$), and the darker the blue the worse is the overestimation ($DN/\sigma^2 < 1$). **c, d** Color maps of $-\log(\lambda_2/N)$ as a function of the side length L of the square domain and the system size N for $\eta = 0.1$ (**c** and **d** correspond to $s = 0.3$ and $s = 3$, respectively). Higher values of $-\log(\lambda_2/N)$ correspond to a lower average connectivity of the system. **e** Scatter plot of the pairs $(-\log(\lambda_2/N), \log(DN/\sigma^2))$, with crosses and circles corresponding to $s = 0.3$ and $s = 3$, respectively. The red lines correspond to the linear fit of the pairs with $-\log(\lambda_2/N)$ lower than 1.50 ($R^2 = 0.08$) and larger than 1.50 ($R^2 = 0.49$), respectively.

In the vast majority of the simulations in which the inference does not match N (especially for large values of L in the high-speed case), we document an underestimation of the size of the collective ($DN/\sigma^2 > 1$). This can be explained by the formation of independent particle clusters¹⁴, which can be revealed from the study of the overall connectivity of the system. As a result, the diffusion coefficient will only be reflective of the size of the cluster to which the focal particle belongs, thereby leading to the observed underestimation. Numerical evidence supporting this explanation is presented in the bottom panels of Fig. 2c, d, where we plot the normalized Fiedler eigenvalue λ_2/N of the Laplacian matrix associated with the average network of interactions during the simulation window, similar to the analysis by Baglietto et al.²⁰.

The Fiedler eigenvalue, measuring the overall connectivity of the system, follows the trend similar to that of DN/σ^2 in Fig. 2a, b. Until a threshold value for the normalized eigenvalue $-\log(\lambda_2/N)$, the method provides a nearly perfect inference of the collective size with a weak dependence on λ_2/N that translates into a modest decrease in accuracy as $-\log(\lambda_2/N)$ increases. On the contrary, beyond the threshold, the accuracy of the inference significantly degrades, with a clear dependence on λ_2/N , as illustrated in Fig. 2e. The origin of this qualitative transition is presently unclear. Perhaps, it is associated with the emergence of ordered bands within the collective²¹, so that a few subgroups are able to synchronize along common headings in a bath of disordered motion for the rest of the collective. We do not favor this explanation, however, given the short length of the

observation window that may challenge the development of such bands. Also, these bands are typically documented for large systems, while the observed threshold phenomenon involves small collectives as well (with N as small as 10).

The analysis of the precision of the inference along with simulations for higher noises ($\eta = 0.3$ and $\eta = 0.5$), different speeds (spanning from 0.1 to 10 at a fine resolution), fewer repetitions (spanning from 100 to 1000), and time steps (spanning from 15 to 150), non-identical initial conditions across repetitions, non-standard Vicsek models (with forward position update), and non-uniform (Gaussian) noise distributions are presented in Supplementary Note 2. Therein, we also include a numerical analysis of the rate of change in time of the interaction network as a function of the particle speed that illustrates the progression from nearly static to memoryless switching topologies, in terms of the decay of an autocorrelation function for the network of interaction.

Theory: low particle speed. We begin with the study of low speed, for which the neighbor set of each particle can be approximated to be constant in time. We hypothesize that the noise is sufficiently small for the system to be in its ordered phase, where particles' headings synchronize. Hence, the collective dynamics of the system is described by the classical noisy consensus problem with static network topology that is extensively studied in control theory²².

Specifically, we consider the following linear dynamics:

$$\boldsymbol{\theta}_{k+1} = \mathbf{A}\boldsymbol{\theta}_k + \mathbf{u}_k, \tag{2a}$$

$$y_k = \boldsymbol{\theta}_k^T \mathbf{e}_f, \tag{2b}$$

where $\boldsymbol{\theta}_k = [\theta_k^1, \dots, \theta_k^N]^T$, \mathbf{A} is the state matrix, $\mathbf{u}_k = [u_k^1, \dots, u_k^N]^T$ is a white noise vector with variance σ^2 (that is, for all k and $\tau \neq 0$: $E[\mathbf{u}_k] = 0$, $E[\mathbf{u}_k \mathbf{u}_k^T] = \sigma^2 \mathbf{I}_N$, and $E[\mathbf{u}_k \mathbf{u}_{k+\tau}^T] = 0$, with $E[\cdot]$ denoting expectation), \mathbf{e}_f is the vector with all zeros but a one in the f th entry, and y_k is the heading of the focal particle. The state matrix \mathbf{A} corresponds to the adjacency matrix of an undirected, unweighted graph with self-loops for each node, such that $\mathbf{A} \mathbf{1}_N = \mathbf{1}_N$ with $\mathbf{1}_N$ being the vector of all ones. The reason why the adjacency matrix is row-stochastic is because the fully ordered phase is a solution of the Vicsek model in the absence of additive noise, so that any common heading of the particles shall be a solution for the collective. Additionally, we hypothesize a connected graph, such that \mathbf{A} is primitive²³.

We prove that the variance of y_k diverges linearly with time, with a slope that is inversely proportional to the system size N . We commence from the dynamics of the covariance matrix $\boldsymbol{\Theta}_k = E[\boldsymbol{\theta}_k \boldsymbol{\theta}_k^T]$ of $\boldsymbol{\theta}_k$ ($\boldsymbol{\Theta}_k$ is the covariance matrix if $E[\boldsymbol{\theta}_0] \mathbf{1}_N = 0$, otherwise, it should be interpreted as the second raw moment),

$$\boldsymbol{\Theta}_{k+1} = E[\boldsymbol{\theta}_{k+1} \boldsymbol{\theta}_{k+1}^T] = \mathbf{A}\boldsymbol{\Theta}_k \mathbf{A}^T + \sigma^2 \mathbf{I}_N, \tag{3}$$

where we used the uncorrelatedness between $\boldsymbol{\theta}_k$ and \mathbf{u}_k , and we introduced the identity matrix \mathbf{I}_N . To solve the linear recursion, we use a similarity matrix \mathbf{V} that is obtained by juxtaposing column-wise the orthonormal eigenvectors of \mathbf{A} . By denoting with $\boldsymbol{\Xi}_k$ the covariance matrix of the transformed state variable $\boldsymbol{\xi}_k = \mathbf{V}^T \boldsymbol{\theta}_k$ and observing that $\boldsymbol{\Xi}_k = \mathbf{V}^T \boldsymbol{\Theta}_k \mathbf{V}$, from Eq. (3) we establish for element ij the following:

$$\boldsymbol{\Xi}_{k+1}^{ij} = \alpha_i \alpha_j \boldsymbol{\Xi}_k^{ij} + \sigma^2 \delta_{ij}, \tag{4}$$

where α_i is the i th eigenvalue of \mathbf{A} , and $\delta_{(i, \cdot)}$ is the Kronecker delta. Since \mathbf{A} is row-stochastic and primitive, 1 is its simple dominant eigenvalue (denoted as α_1) whose eigenvector is $\mathbf{v}_1 = \mathbf{1}_N / \sqrt{N}$ ²³. As $k \rightarrow \infty$, the off-diagonal elements of $\boldsymbol{\Xi}_k$ converge to zero, and the i -th diagonal entry, for $i \neq 1$, converges

to $\sigma^2 / (1 - \alpha_i^2)$. The first diagonal entry, instead, diverges linearly with time, according to the expression $\boldsymbol{\Xi}_k^{11} = \boldsymbol{\Xi}_0^{11} + \sigma^2 k$.

Given Eq. (2b), the variance of y_k is $Y_k = \mathbf{e}_f^T \boldsymbol{\Theta}_k \mathbf{e}_f$. Working with the transformed state variable $\boldsymbol{\xi}_k$, we have $Y_k = \mathbf{e}_f^T \mathbf{V} \boldsymbol{\Xi}_k \mathbf{V}^T \mathbf{e}_f$, which, for sufficiently large k , yields the sought dependence on the size of the collective

$$Y_k \approx \frac{\sigma^2}{N} k. \tag{5}$$

Here, we used the facts that the first column of \mathbf{V} is $\mathbf{1}_N / \sqrt{N}$ and that the first diagonal entry of $\boldsymbol{\Xi}_k$ diverges in time with a rate of σ^2 . Similar to the mean square displacement of a particle in a gas¹⁵, Y_k linearly diverges and the diffusion coefficient is $D = \sigma^2 / N$, thereby offering mathematical backing to the numerical results in Fig. 2a ($DN / \sigma^2 \approx 1$). In particular, the values of the diffusion coefficient and variance of the added noise are sufficient to accurately infer the collective size under the approximation of a static topology. This result may be counterintuitive if one considers that for an XY-model interactions are completely local, so that intuition may suggest that the trajectory of a particle only contains the footprint of the dynamics of its neighbors. However, this is not the case since the connectivity of the interaction network ensures that, albeit indirectly, every particle in the collective interacts with any other particle.

A few comments are in order for the derivation of Eq. (5). First, we considered a more general form of additive noise in Eq. (2a) than the Vicsek model, to emphasize that only knowledge of the variance is sufficient for the inference and that other noise distributions can be equivalently considered. This evidence strengthens the connection between the notion of thermodynamic temperature and the variance of the noise^{16–19}. Second, should the network be disconnected as in the case of high values of L , the same mathematical steps would yield a higher diffusion coefficient of σ^2 / N_f , where N_f is the number of particles in the connected component the focal particle belongs to. Hence, the proposed inference approach would never overestimate the size of the system, see Fig. 2. Third, the approach is robust with respect to individual variations in the form of heterogeneous variances $\sigma_1^2, \dots, \sigma_N^2$ across the particles. In this case, the same derivations would yield an equivalent linear dependence in time for Y_k with $D = \sum_{i=1}^N \sigma_i^2 / N$, so that the inference is independent of the choice of the focal particle and the only required information is the average value of the variance. Likewise, the approach is robust with respect to weighted interactions, whereby the results would not change for any undirected primitive row-stochastic matrix \mathbf{A} .

Finally, although the derivation of Eq. (2b) assumes the network to be constant in time and given once for all across the repetitions, its applicability is likely to extend to a more general setting that should be formalized in future studies. Specifically, should one consider slowly-varying topologies so that \mathbf{A} varies in time much slower than the rate of divergence of the variance of y_k , the same mathematical arguments might be applicable. Likewise, should the adjacency matrix differ across repetitions, we would not expect changes in Eq. (2b), provided each of the matrices is connected, since no topological feature of the adjacency matrix enters the estimation of the size of the collective. Simulation results in Fig. 2 and in Supplementary Note 2 offer compelling evidence for these claims, whereby the inference is accurate albeit the network topology changes in time due to additive noise, and initial conditions are varied across repetitions.

Theory: high particle speed. The analytical result in Eq. (5) is based on the assumption of a static topology. An equivalent claim can be established in the opposite scenario of random mixing,

where the speed of the particles is sufficiently large for the memory of past interactions to fade between two consecutive time steps. This limit case can be studied through the vectorial network model (VNM)²⁴, which implements heading update by randomly assigning to any particle K neighbors drawn uniformly from the entire collective at each time step. In this high-speed limit, particles interact on average with $\pi/L^2(N-1)$ other particles within their unit circles so that L and K can be related by $L^2 = \pi(N-1)/(K-1)$. For the case of small noise, the model reduces to²⁵

$$\begin{aligned}\boldsymbol{\theta}_{k+1} &= \mathbf{A}_k \boldsymbol{\theta}_k + \mathbf{u}_k, \\ \gamma_k &= \boldsymbol{\theta}_k^T \mathbf{e}_f,\end{aligned}\quad (6)$$

where the matrices $\mathbf{A}_0, \mathbf{A}_1, \dots$ are independent identically distributed (i.i.d.) with common random variable \mathbf{A} , and the noise \mathbf{u}_k and all the other variables are defined as in the case of static topology. Specifically, \mathbf{A} is such that each of its rows is the sum of $K \geq 2$ i.i.d. vectors, each one having all entries 0, except one (selected from a uniform distribution) that is equal to $1/K$. Similar to the low-speed case in Eq. (2a) and consistent with the original Vicsek model (1), matrix \mathbf{A} is row-stochastic so that $\mathbf{A}\mathbf{1}_N = \mathbf{1}_N$ by construction. This temporal patterning yields an all-to-all average network, $E[\mathbf{A}] = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$. We acknowledge the possibility that the number of distinct neighbors is less than K within the VNM; however, the probability that two vectors will contribute a non-zero entry at the same location is very small.

Similar to Eq. (3), we write

$$\boldsymbol{\Theta}_{k+1} = E[\mathbf{A}_k \boldsymbol{\theta}_k \boldsymbol{\theta}_k^T \mathbf{A}_k^T] + \sigma^2 \mathbf{I}_N \quad (7)$$

but, in this case, the adjacency matrix depends on time and cannot be moved out of the expectation. To establish a linear recursion for $\boldsymbol{\theta}_k$, we resort to Kronecker algebra,

$$\text{vec}(\boldsymbol{\Theta}_{k+1}) = \mathbf{G}\text{vec}(\boldsymbol{\Theta}_k) + \sigma^2 \text{vec}(\mathbf{I}_N). \quad (8)$$

Here, “vec” identifies the process of vectorization and $\mathbf{G} = E[\mathbf{A} \otimes \mathbf{A}]$ can be explicitly computed as²⁵

$$\mathbf{G} = \frac{1}{KN} \text{vec}(\mathbf{I}_N) \text{vec}(\mathbf{R})^T + \frac{1}{N^2} \mathbf{1}_N \mathbf{1}_N^T \otimes \mathbf{1}_N \mathbf{1}_N^T, \quad (9)$$

with “ \otimes ” being the outer product and $\mathbf{R} = \mathbf{I}_N - \mathbf{1}_N \mathbf{1}_N^T / N$ a projection matrix. None of the analytical developments by Porfiri²⁵ can be utilized for the case at hand, whereby the dynamics considered therein was deprived of the mean to study the evolution of the synchronization error. In Supplementary Note 3, we study the spectral properties of \mathbf{G} to determine that, for sufficiently large times, element ij of $\boldsymbol{\Theta}_k$ can be written as

$$\Theta_k^{ij} \approx \frac{\sigma^2 \gamma}{N} k, \quad (10)$$

where $\gamma = (a-b)/(a-N+1)$, with $a = N(N-1-2KN+2K+K^2N)$ and $b = N(N-1-KN+K)$. Given that $Y_k = \Theta_k^{ff}$, we find the sought dependence on the size of the collective,

$$Y_k \approx \frac{\sigma^2 \gamma}{N} k. \quad (11)$$

Different from the case of low speed, the particle density enters the estimation of the diffusion coefficient through γ . As further detailed in Supplementary Note 3, for any choice of $2 \leq K \leq N$ and $N > 1$, γ is between 1 and 2. The largest values of γ are registered for low densities, that is, small systems in large domains. Increasing the density (that is, increasing K) lowers the value of γ , thereby leading to accurate inferences. This theoretical result is in qualitative agreement with numerical simulations in Fig. 2b, which indicate that in the high-speed case, the accuracy of the inference degrades with L and improves with N . However, care should be placed when drawing conclusions for large values of L that lead to poorly

connected average networks (as evidenced by the normalized Fiedler eigenvalue in Fig. 2d), in contrast with the all-to-all average network predicted by the vectorial network model.

In principle, the computation can be extended to weighted, random interactions, in which every node randomly weights the state of its K neighbors at each time-step, such as blinking models²⁶ activity-driven networks²⁷, and averaging models²⁸. However, this would likely lead to a different expression for γ since the spectral properties of \mathbf{G} would vary. The presence of heterogeneities in the added noise across the units will not produce a difference in the diffusion coefficient with respect to the static case, as shown in Supplementary Note 4.

Conclusions

There is a growing effort to build physical intuition and technical rigor in collective dynamics by recognizing analogies with thermodynamics^{29–33}. For example, Sinhuber and Oulette³⁰ have introduced definitions of pressure and chemical potentials for insect swarms to study the coexistence of a condensed phase surrounded by a vapor phase observed in experiments. Likewise, Haeri et al.³² have recently elucidated thermodynamic variables of collectives governed by attractive and repulsive forces, presenting an equation of state similar to the ideal gas law.

Here, we unveil a further, microscopic analogy, upon which we establish a methodology to infer the size of a collective from the dynamics of a single unit. Different from the entire body of knowledge that relies on deterministic dynamics^{6–9}, we treat stochasticity as a commodity to solve the inference problem. Just as the motion of a Brownian particle in a liquid can be used to determine Avogadro’s number as shown in the classical experiments by Perrin¹², so we demonstrate that the stochastic motion of a single unit is sufficient for the determination of the size of the entire collective.

We focus on the ordered (homogeneous) phase of the self-propelled particle model of Vicsek, for which we discover an inverse proportionality between the diffusion coefficient associated with the heading of any of the particles and the size of the collective. Given the sole knowledge of the temperature of the system, in the form of the variance of the added noise, one can successfully infer the size of the collective from the diffusion coefficient. We offer detailed analytical treatment of two limit cases of the Vicsek model, encompassing the XY -model and mean-field behavior of a ferromagnet. The mathematical study of these cases confirms numerical predictions that connectivity of the collective ensures that the diffusion coefficient is, within a first approximation, independent of the particle density—a counter-intuitive finding with respect to the case of an ideal gas where the geometry of the system modulates the diffusion coefficient¹⁵. While these limited cases may offer partial insight into the study of key nonlinear features of the Vicsek model, like the nature of its phase transition²¹, they accurately anticipate the observed evolution of the variance of the heading and its dependence on the system parameters at low temperatures.

This study is not free of limitations, which shall be addressed in future work toward the study of real collectives, from insect swarms to bird flocks, fish schools, and human crowds. First, the approach is demonstrated on a rather simplified model of collective behavior, which, albeit popular and general, does not capture all the intricacies and complexities of collective behavior. Future work should seek to apply the approach to more detailed models of collective behavior, including, for example, attraction and repulsion rules². Second, the implementation of the approach requires knowledge of the noise variance, which may not be feasible all the times, especially when dealing with experimental observations. Should one have access to some video recordings of the collective and knowledge of an adequate mathematical model, it could be possible to

infer the value of the noise through dimensionality reduction techniques, without the need for tracking³⁴.

Overall, this study provides a rigorous, mathematically backed method to infer the size of a realistic collective from measurements of some of its units, whose random motion contains the footprints of the entire system. The theoretical underpinnings of the method provide further evidence for the analogies identified by Einstein between interdisciplinary research in the collective behavior of animal groups and modern physics³⁵.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code that supports the findings of this study is available from the corresponding author upon reasonable request.

Received: 15 November 2021; Accepted: 18 March 2022;

Published online: 11 April 2022

References

- Sumpter, D. J. The principles of collective animal behaviour. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 5–22 (2006).
- Vicsek, T. & Zafeiris, A. Collective motion. *Phys. Rep.* **517**, 71–140 (2012).
- Pilkiewicz, K. et al. Decoding collective communications using information theory tools. *J. R. Soc. Interface* **17**, 20190563 (2020).
- Strandburg-Peshkin, A., Papageorgiou, D., Crofoot, M. C. & Farine, D. R. Inferring influence and leadership in moving animal groups. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170006 (2018).
- Flack, A., Nagy, M., Fiedler, W., Couzin, I. D. & Wikelski, M. From local collective behavior to global migratory patterns in white storks. *Science* **360**, 911–914 (2018).
- Haehne, H., Casadiego, J., Peinke, J. & Timme, M. Detecting hidden units and network size from perceptible dynamics. *Phys. Rev. Lett.* **122**, 158301 (2019).
- Porfiri, M. Validity and limitations of the detection matrix to determine hidden units and network size from perceptible dynamics. *Phys. Rev. Lett.* **124**, 168301 (2020).
- Tang, X. et al. Dynamical network size estimation from local observations. *N. J. Phys.* **22**, 093031 (2020).
- Tyloo, M. & Delabays, R. System size identification from sinusoidal probing in diffusive complex networks. *J. Phys. Complex.* **2**, 025016 (2021).
- Einstein, A. Zur theorie der brownischen bewegung. *Ann. Phys.* **324**, 371–381 (1906).
- Smoluchowski, V. M. V. & im unbegrenzten Raum, I. D. Zusammenfassende bearbeitungen. *Ann. Phys.* **21**, 756 (1906).
- Perrin, J. Mouvement brownien et réalité moléculaire. *Ann. Chim. Phys.* **18**, 5–114 (1909).
- Feynman, R. P., Leighton, R. B. & Sands, M. *The Feynman Lectures On Physics, Vol. I: The New Millennium Edition: Mainly Mechanics, Radiation, and Heat, Vol. 1* (Basic Books, 2011).
- Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I. & Shochet, O. Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**, 1226 (1995).
- Cussler, E. L. *Diffusion: Mass Transfer in Fluid Systems* (Cambridge University Press, 2009).
- Czirók, A., Vicsek, M. & Vicsek, T. Collective motion of organisms in three dimensions. *Phys. A Stat. Mech. Appl.* **264**, 299–304 (1999).
- Vicsek, T., Czirók, A., Farkas, I. J. & Helbing, D. Application of statistical mechanics to collective motion in biology. *Phys. A: Stat. Mech. Appl.* **274**, 182–189 (1999).
- Porfiri, M. & Ariel, G. On effective temperature in network models of collective behavior. *Chaos* **26**, 043109 (2016).
- Vicsek, T. *Fluctuations and Scaling in Biology* (Oxford University Press, 2001).
- Baglietto, G., Albano, E. V. & Candia, J. Complex network structure of flocks in the standard Vicsek model. *J. Stat. Phys.* **153**, 270–288 (2013).
- Ginelli, F. The physics of the Vicsek model. *Eur. Phys. J. Spec. Top.* **225**, 2099–2117 (2016).
- Olfati-Saber, R., Fax, J. A. & Murray, R. M. Consensus and cooperation in networked multi-agent systems. *Proc. IEEE* **95**, 215–233 (2007).
- Horn, R. A. & Johnson, C. R. *Matrix Analysis* (Cambridge University Press, 2012).
- Aldana, M., Dossetti, V., Huepe, C., Kenkre, V. M. & Larralde, H. Phase transitions in systems of self-propelled agents and related network models. *Phys. Rev. Lett.* **98**, 095702 (2007).
- Porfiri, M. Linear analysis of the vectorial network model. *IEEE Trans. Circuits Syst. II Express Briefs* **61**, 44–48 (2014).
- Belykh, I. V., Belykh, V. N. & Hasler, M. Blinking model and synchronization in small-world networks with a time-varying coupling. *Phys. D Nonlinear Phenom.* **195**, 188–206 (2004).
- Liu, S., Perra, N., Karsai, M. & Vespignani, A. Controlling contagion processes in activity driven networks. *Phys. Rev. Lett.* **112**, 118702 (2014).
- Cao, F. *k*-averaging agent-based model: propagation of chaos and convergence to equilibrium. *J. Stat. Phys.* **184**, 1–19 (2021).
- Giannini, J. A. & Puckett, J. G. Testing a thermodynamic approach to collective animal behavior in laboratory fish schools. *Phys. Rev. E* **101**, 062605 (2020).
- Sinhuber, M. & Ouellette, N. T. Phase coexistence in insect swarms. *Phys. Rev. Lett.* **119**, 178003 (2017).
- Sinhuber, M., van der Vaart, K., Feng, Y., Reynolds, A. M. & Ouellette, N. T. An equation of state for insect swarms. *Sci. Rep.* **11**, 1–8 (2021).
- Haeri, H., Jerath, K. & Leachman, J. Thermodynamics-inspired macroscopic states of bounded swarms. *Lett. Dynamic Syst. Control* **1**, 011015 (2021).
- Crosato, E., Spinney, R. E., Nigmatullin, R., Lizier, J. T. & Prokopenko, M. Thermodynamics and computation during collective motion near criticality. *Phys. Rev. E* **97**, 012120 (2018).
- Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V. & Scotti, F. A survey of unsupervised generative models for exploratory data analysis and representation learning. *ACM Comput. Surv.* **54**, 1–40 (2021).
- Dyer, A. G. et al. Einstein, von Frisch and the honeybee: a historical letter comes to light. *J. Compar. Physiol. A.* **207**, 449–456 (2021).

Acknowledgements

The authors wish to thank Agnieszka Truszkowska for setting up and performing numerical simulations for large collectives. P.D.L. was supported by the program “STAR 2018” of the University of Naples Federico II and Compagnia di San Paolo, Istituto Banco di Napoli - Fondazione, project ACROSS. M.P. was supported by the National Science Foundation under grant numbers CMMI 1561134 and CMMI 1932187.

Author contributions

P.D.L. and M.P. designed the study, performed the numerical analysis, developed the theoretical framework, analyzed and interpreted the results, and wrote the final draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42005-022-00864-9>.

Correspondence and requests for materials should be addressed to Pietro De Lellis or Maurizio Porfiri.

Peer review information *Communications Physics* thanks Joseph Lizier, Robin Delabays, Melvyn Tyloo, and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022