Intrinsic Dimension Estimation for Discrete Metrics

Iuri Macocco,[1] Aldo Glielmo,[1,2] Jacopo Grilli,[3] and Alessandro Laio[1,3,*]
[1] *International School for Advanced Studies (SISSA), Via Bonomea 265, 34136 Trieste, Italy*
[2] *Bank of Italy, DG for Information Technology, 00044 Rome, Italy*
[3] *The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34014 Trieste, Italy*

(Received 13 May 2022; revised 16 September 2022; accepted 20 January 2023; published 8 February 2023)

Real-world datasets characterized by discrete features are ubiquitous: from categorical surveys to clinical questionnaires, from unweighted networks to DNA sequences. Nevertheless, the most common unsupervised dimensional reduction methods are designed for continuous spaces, and their use for discrete spaces can lead to errors and biases. In this Letter we introduce an algorithm to infer the intrinsic dimension (ID) of datasets embedded in discrete spaces. We demonstrate its accuracy on benchmark datasets, and we apply it to analyze a metagenomic dataset for species fingerprinting, finding a surprisingly small ID, of order 2. This suggests that evolutive pressure acts on a low-dimensional manifold despite the high dimensionality of sequences' space.

DOI: 10.1103/PhysRevLett.130.067401

Data produced by experiments and observations are very often high dimensional, with each data point being defined by a sizeable number of features. To the pleasure of modelers, real-world datasets seldom occupy this high-dimensional space uniformly, as strong regularities and constraints emerge. Such a property is what allows for low-dimensional descriptions of these high-dimensional data, ultimately making science possible.

In particular, data points are often effectively contained in a manifold which can be described by a relatively small number of coordinates. The number of such coordinates is called intrinsic dimension (ID). More formally, the ID is defined as the minimum number of variables needed to describe the data without significant information loss. Its knowledge is of paramount importance in unsupervised learning [1–3] and has found applications across disciplines. In solid-state physics and statistical physics, the ID can be used as a proxy of an order parameter describing phase transitions [4,5]; in molecular dynamics it can be used to quantify the complexity of a trajectory [6]; in deep learning theory the ID indicates how information is compressed throughout the various layers of a network [7–9]. During the last three decades much progress has been made in the development of sophisticated tools to estimate the ID [10,11], and most estimators have been formulated (and are supposed to work) in spaces where distances can vary continuously. However, many datasets are characterized by discrete features and, consequently, discrete distances. For instance, categorical datasets like satisfaction questionnaires, clinical trials, unweighted networks, spin systems, protein, and DNA sequences fall into this category.

Two main methods are usually employed in these cases. The box counting (BC) estimator [12–14]—which is defined by measuring the scaling between the number of boxes needed to cover a dataset and the boxes' size—provides good results for two-dimensional to three-dimensional datasets but is computationally demanding for higher-dimensional datasets. The second popular method is the fractal dimension (FD) estimator [12,15,16], and it is based on the assumption of a power law relationship $N \sim r^d$ for the number N of neighbors within a sphere of radius r from a given point, where d is the fractal dimension of the data. This estimator has been successfully applied, on discrete datasets, to model the phenomena of dielectric breakdown [17] and Anderson localization [18]. For nonfractal objects, both methods are reliable only in the limit of small boxes and small radii, since the manifold containing the data can be curved, and the data points can be distributed nonuniformly [19]. However, in discrete spaces such a limit is not well defined due to the minimum distance induced by any discrete lattice, and this can lead to systematic errors [20,21].

In this Letter, we introduce an ID estimator explicitly formulated for spaces with discrete features. In discrete spaces, the ID can be thought of as the dimension of a (hyper)cubic lattice where the original data points can be (locally) projected without a significant information loss. The key challenge in dealing with the discrete nature of the data lies in the proper definition of volumes on lattices. To this end, we introduce a novel method that makes use of Ehrhart's theory of polytopes [22], which allows one to enumerate the lattice points of a given region. By measuring a suitable statistics, depending on the number of data points observed within a given (discrete) distance, one can infer the value of the dimension of the region, which we interpret as the ID of the dataset. The statistics we use is defined in such a way that density of points is required to be constant only locally and not in the whole dataset.

Importantly, our estimator allows one to explicitly select the scale at which the ID is computed.

Methods.—We assume data points to be uniformly distributed on a generic domain, and that their density is ρ. In such domain, we consider a region A with volume $V(A)$. Since we are assuming points to be independently generated, the probability of observing n points in A is given by the Poisson distribution [23]

$$P(n, A) = \frac{[\rho V(A)]^n}{n!} e^{-\rho V(A)} \tag{1}$$

so that $\langle n \rangle = \rho V(A)$. Consider now a data point i and two regions A and B, one containing the other, and both containing the data point $i \in A \subset B$. Then the number of points n and $k - n$ falling, respectively, in A and $B \backslash A$ are Poisson distributed with rates $\lambda_1 = \rho V(A)$ and $\lambda_2 = \rho V(B \backslash A)$. The conditional probability of having n points in A given that there are k points in B is

$$P(n|k) = \frac{P(n)P(k-n)}{P(k)} = \binom{k}{n} p^n (1-p)^{k-n} \tag{2}$$

with

$$p = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\rho V(A)}{\rho V(B)} = \frac{V(A)}{V(B)}. \tag{3}$$

Thus $n|k \sim \text{Binomial}(n; k, p)$. As far as the density ρ is constant within A and B, p is simply equal to the ratio of the volumes of the considered regions and, remarkably, density independent. This is a key property which, as we will show, allows for using the estimator even when the density is approximately constant only locally, and varies, even substantially, across larger distance scales. One can then write a conditional probability of the observations n_i (one for each data point), given the parameters k_i and p_i, which can possibly be point dependent:

$$\mathcal{L}(n_i|k_i, p_i) = \prod_{i=1}^{N} \text{Binomial}(n_i|k_i, p_i). \tag{4}$$

Such formulation assumes all the observations to be statistically independent. Strictly speaking this is typically not true, since the regions A and B of different points can be overlapping. We will address this issue in the Supplemental Material [24], demonstrating that neglecting correlations does not induce significant errors.

The next step consists of defining the volumes in Eq. (3) according to the nature of the embedding manifold. We now assume our space to be a lattice where the L^1 metric is a natural choice. In this space the volume $V(A)$ is the number of lattice points contained in A. According to Ehrhart theory of polytopes [25], the number of lattice points within distance t in dimension d from a given point amounts to [26]

$$V_\diamond(t, d) = \binom{d+t}{d} {}_2F_1(-d, -t, -d-t, -1) \tag{5}$$

where ${}_2F_1(a, b, c, z)$ is the ordinary hypergeometric function. At a given t, the above expression is a polynomial in d of order t. As a consequence, the ratio of volumes defining the value of p in Eq. (3) becomes a ratio of two polynomials in d. Given a dataset, the choice of t_1 and t_2 fixes the values of n_i and k_i in the expression for the likelihood. The maximization of the likelihood function [Eq. (4)] with respect to d allows one to infer the data manifold's ID, which is simply given the root of equation (see the Supplemental Material [24] for more details on the derivation)

$$\frac{V_\diamond(t_1, d)}{V_\diamond(t_2, d)} - \frac{\langle n \rangle}{\langle k \rangle} = 0 \tag{6}$$

where the mean value over n and k is intended over all the points of the dataset. The root can be easily found with standard optimization libraries. This procedure defines an ID estimator that, for brevity, we will call I3D (intrinsic dimension estimator for discrete datasets).

Very importantly, the ID estimate is density independent as such a factor cancels out [see Eq. (3)]. The error on the estimator has a theoretical lower bound, given by the Cramer-Rao inequality, which has an explicit analytic expression. As an alternative, the ID can be estimated by a Bayesian approach as the mean value of its posterior distribution, and the error estimated via the posterior variance (details in the Supplemental Material [24]).

The estimation of the ID depends on the choice of the volumes of the smaller and larger regions, which are parametrized by the "radii" t_1 and t_2. By varying t_2, the radius of the largest probe region, one can explore the behavior of the ID at different scales. The proper range of t_2 is dataset dependent and should be chosen by plotting the value of the ID as a function of it, as we will illustrate in the following. If the dataset has a well-defined ID, one will observe a (approximate) plateau in this plot. This leaves the procedure with one free parameter: the ratio $r = t_1/t_2$ and its choice influence the statistical error. In continuous space the ratio between volumes in Eq. (3) is simply $p = r^d$, and the Cramer-Rao variance has a simple dependence on the parameter r. By minimizing it with respect to r, one obtains that the optimal value for the ratio is $r_{\text{opt}} \sim 0.2032^{(1/d)}$ (see the Supplemental Material [24]).

In order to check the goodness of the estimator, we test whether the number of points n contained within the internal shells is actually distributed as a mixture of binomials, as our model assumes:

$$P(n) = \sum_k P(k) \mathrm{B}\left(n; k, \frac{V_\diamond(t_1, d)}{V_\diamond(t_2, d)}\right) \tag{7}$$

where $P(k)$ is the empirical probability distribution of k found by fixing t_2. In the following we will compare the

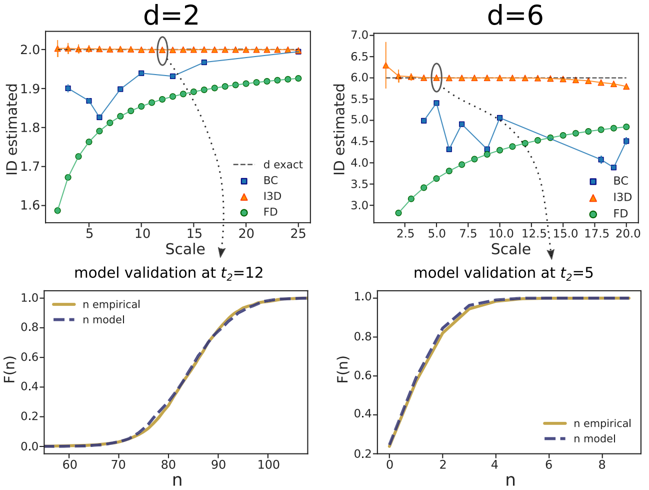

FIG. 1. Performance of I3D, BC, and FD estimators for points uniformly distributed on a square lattice of size 50 in 2D and size 20 in 6D. Datasets were obtained by sampling, respectively, 20 realizations of 2500 and 100 000 points. Error bars are given by the standard deviation over the different realizations. Lower panels: I3D model validation performed by comparing empirical and theoretical cdfs of the random variable n.

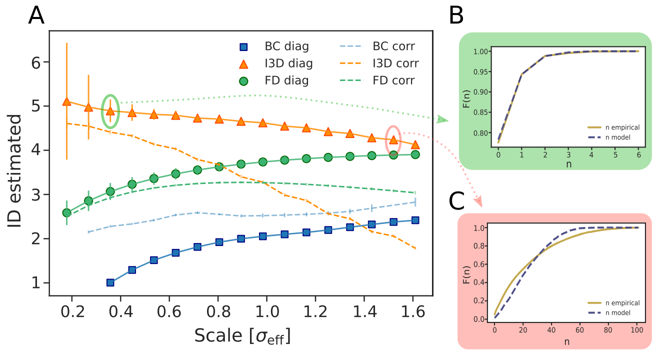

FIG. 2. ID estimations of I3D, BC, and FD on 20 realizations of 2500 points drawn from a Gaussian distribution in 5D and projected on a lattice (a). Solid lines with markers are related to diagonal covariance matrix, dashed lines to the nondiagonal case. Panels (b) and (c) show, respectively, I3D model validation at small and large scales.

empirical cumulative distribution of n to the cumulative distribution of $P(n)$.

Results: Uniform distribution.—We tested the I3D estimator on artificial datasets, and compared it against the two aforementioned methods: the box counting and the fractal dimension (FD). The BC estimate of the ID is obtained by a linear fit between the logarithm of the number of occupied covering boxes and the logarithm of the boxes' side. Seemingly, for the FD, the linear fit is computed among the logarithm of the average number of neighbors within a given radius and the logarithm of the radius. In both cases, the scale reported in the figures is given by the largest box or radius included in the fit. We started by analyzing uniformly distributed points in 2D and 6D square lattices. We adopted periodic boundary conditions in order to reduce boundary effects as much as possible. For the I3D estimator, in this and all following cases, we set $t_1/t_2 = r = 0.5$. The results are shown in Fig. 1. While the BC and FD proved to be reliable in finding the fractal dimension of repeating, self-similar lattices [12,17], they do not manage to assess the proper dimension of randomly distributed points, especially at small scales. The I3D estimator, instead, returns accurate values for the ID at all scales and, importantly, provides the correct estimate also on self-similar lattices (see the Supplemental Material [24]). Remarkably the I3D estimator allows one to select the scale explicitly by varying the radius t_2. In the lower panels of Fig. 1, we also report a first example of model validation for I3D. The two cumulative distribution functions [empirical and theoretical one, according to Eq. (7)] perfectly match, meaning that the ID estimation is reliable.

Gaussian distribution.—Secondly, we tested the estimators on Gaussian distributed points in five dimensions, analyzing a case in which the data are uncorrelated and a case in which a correlation is induced by a nondiagonal covariance matrix. In both cases, we set diagonal elements of the covariance matrix to $\sigma = 5$ (implying an effective standard deviation of the distribution of $\sigma_{\rm eff} = \sqrt{d}\sigma$), while off diagonal terms—for correlated data—were uniformly extracted in the interval (0,2). The values were chosen in order to keep the dimension of the dataset under control, as correlations of the same order of the diagonal would reduce the dimensionality of the dataset. The points were projected on a lattice by taking the nearest integer in each coordinate. As one can observe in panel (a) of Fig. 2, I3D is accurate as far as it explores a neighborhood where the density does not vary too much (namely, as far as $t_2/\sigma_{\rm eff} \lesssim 1$). Correspondingly, empirical and model cdfs in panel (b) are superimposed. Beyond such distance, neighborhoods are characterized by nonconstant density; consequently, estimates get less precise and, accordingly, the two cdfs show inconsistencies (panel (c): $t_2/\sigma_{\rm eff} \sim 1.5$). On the other hand, the BC and FD estimations are far from desired values at any scales, for both correlated and uncorrelated cases.

Spin dataset.—As a third test, we created synthetic Ising-like spin systems with a tunable ID, which is given by the number of independent parameters used to generate the dataset. The 1D ensemble is obtained by generating a set of points belonging to a line embedded in $\mathbb{R}^D$ with the process $\boldsymbol{\varphi}_i = \boldsymbol{\varphi}_0 + \boldsymbol{\alpha}\epsilon(i)$. Here, $\boldsymbol{\alpha}$ is a fixed random vector of unitary norm with uniformly distributed components, and $\boldsymbol{\varphi}_0 = -0.5$ is the y intercept that, for simplicity, is equal for all the components; ϵ_i are Gaussian distributed, $\epsilon \sim \mathcal{N}(0, 10)$, and independently drawn for each sample i. We then proceed to the discretization by extracting the $z_i = \text{sign}(\boldsymbol{\varphi}_i)$, an ensemble of N states of D discrete spins. The pipeline is summarized in Fig. 3. The role of $\boldsymbol{\varphi}_0$ is to introduce an offset in order to enhance the number of the

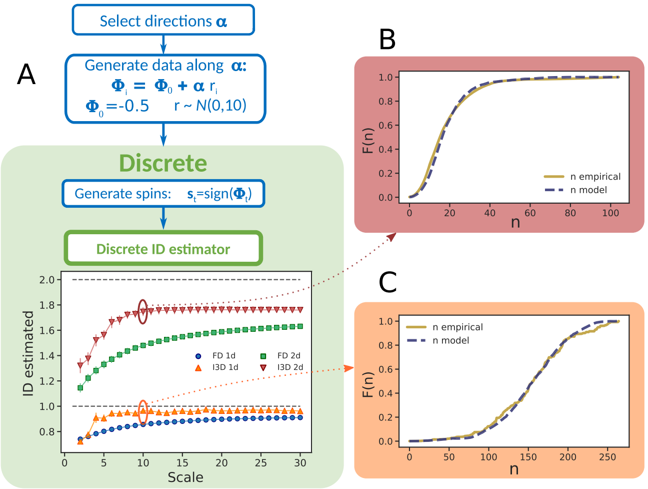

FIG. 3. (a) The pipeline used to create an ensemble of binary spins with a low ID, together with the results of FD and I3D estimators on 1D and 2D datasets. I3D estimations were validated by comparing theoretical and empirical cdfs [panels (b) and (c)].

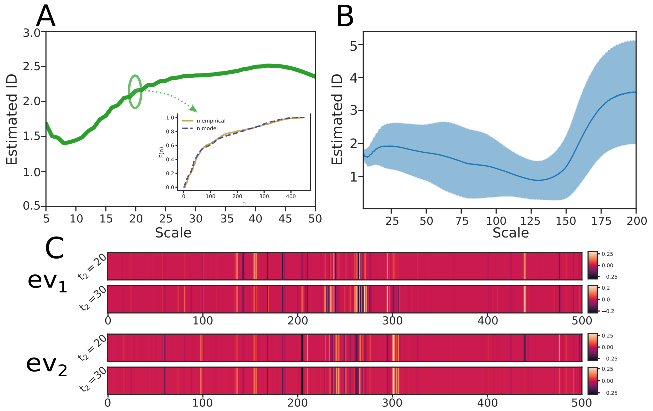

FIG. 4. Estimated ID at small to medium distances for one of the clusters of the genomics dataset [panel (a)]. The inset reports the fair superposition of empirical and modeled cdfs of n. Panel (b) shows average and standard deviation of the IDs estimated separately for each cluster. Panel (c) shows first and second principal component analysis (PCA) eigenvectors of the data points within given distances t_2 (20 or 30) from the center of the cluster used for panel (a).

reachable discrete states. In fact, for $\boldsymbol{\varphi}_0 = 0$, we would obtain only two different states, given by $\boldsymbol{z} = \mathrm{sign}(\boldsymbol{\alpha}\epsilon) = \pm\mathrm{sign}(\boldsymbol{\alpha})$, since the spins would change sign synchronously. An offset $\neq 0$ allows the angles $\boldsymbol{\varphi}_i$ and the spins $\boldsymbol{z}_i$ to shift sign in an asynchronous way. The extension to higher dimensions is straightforward and consists of generating the initial points as $\boldsymbol{\varphi}_i = \boldsymbol{\varphi}_0 + \sum_{j=1}^{id} \boldsymbol{\alpha_j}\epsilon_j(i)$, with $\boldsymbol{\alpha_j} \cdot \boldsymbol{\alpha_k} \sim \delta_{jk}$. Because of the nature of data domain (a D-dimensional hypercube with side 1), the BC cannot be applied, as boxes with side larger than 1 would include the whole dataset. FD and I3D estimates for the 1D system are very close. This is not surprising as both continuous and discrete volumes (and, consequently, the neighbors) scale linearly with the radius. In the 2D case, I3D clearly outperforms the other methods, although even the best estimate remains slightly lower than the true value. This effect, due to nonuniform density, is relatively small, and indeed the empirical and theoretical cdfs are rather consistent [panels(b) and (c)]. Such an effect becomes more important as the dimension rises (see the Supplemental Material [24] for examples in $d = 3$ and $d = 4$).

16S Genomics strands.—Lastly, we present the application of our methodology to a real-world dataset in the field of genomics. The dataset consists of DNA sequences of ~100–300 nucleotides. We selected a dataset downloaded from the Qiita server ([27]) [28]. In such a study, they sequenced the $v4$ region of the 16S ribosomal RNA of the microbiome associated with sponges and algal water blooms. This small subunit of rRNA genes is widely used to study the composition of microbial communities [29–32]. Hamming distance and the binary mapping $A\!:\!11, T\!:\!00, C\!:\!10$, and $G\!:\!01$ were used to compute sequences' distance. The canonical letter representation leads to almost identical results (see the Supplemental Material [24]). To avoid dealing with isolated sequences, we kept only sequences having at least ten neighbors within a distance of 10. Sequences come with their associated multiplicity, related to the number of times the same read has been found in the samples. We ignore such degeneracy and compute an ID which describes just the distribution of the points regardless of their abundance.

To begin with, we estimated the ID on a subset of sequences that are similar to each other. In order to find such sets, we perform a k-means clustering and calculate the ID separately for each of them. Panel (a) in Fig. 4 shows the ID at small to medium scale for one such cluster. The empirical and reconstructed cdfs, performed at $t_2 = 20$ (see inset), are fairly compatible. Panel (b) shows the average and the standard deviation of the ID of all clusters (weighted according to the respective populations). One can appreciate that the ID is always between 1 and 3 in a wide range of distances, showing a plateau around 2 for $15 < t_2 < 40$.

Such a low value for the ID is an interesting and unexpected feature, as it suggests that, despite the high-dimensionality of sequences' space, evolution effectively operates in a low-dimensional space. Qualitatively, an ID ~ 2 on a scale of ~20 means that if one considers all the sequences differing by approximately 20 mutations from a given sequence, these mutations cannot be regarded as independent one from each other, but are correlated in such a way that approximately 18 degrees of freedom are effectively forbidden. The "direction" of these correlated mutations can be, at least approximately, measured by performing PCA in the space of sequences with the binary mapping. The first two dominant eigenvectors, shown in panel (c), were estimated using all the sequences within a distance of 20 (top) and 30 (bottom) from the center of the cluster of panel (a). Remarkably, the eigenvectors do not change significantly on this distance range, indicating that, consistently with the low value of the ID, the data manifold on this scale can be approximately described by a two-dimensional plane. In order to provide an interpretation of the vectors defining this plane, we repeated this same analysis on the previously mentioned spin model. In this case, if the generative model is defined by two vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, the first two dominant eigenvectors of a PCA performed on ∼1000 points are contained in the span of the two generating vectors, with a residual of 0.04 (see the Supplemental Material [24] for details). The components of a vector $\boldsymbol{\alpha}$ can then be qualitatively interpreted as proportional to the mutation probabilities of the associated nucleotide for a collective mutation process. In the genomics dataset this reasoning can applied only locally: the direction of correlated mutation is significantly different in different clusters, indicating that the data manifold is highly curved.

Conclusions.—We presented an ID estimator formulated to analyze discrete datasets. Our method relies on few mathematical hypotheses and is asymptotically correct if the density is constant within the probe radius t_2. In order to prove the estimator's effectiveness, we tested the algorithm against three different artificial datasets and compared it to the well known box counting and fractal dimension estimators. While the last two performed poorly, the new one achieved good results in all cases, providing reliable ID estimations corroborated by the comparison of empirical and model cumulative distribution functions for one of the observables. We finally applied the estimator on a genomics dataset, finding an unexpectedly low ID which hints at strong constraints in the sequences' space, and then exploited such information to give a qualitative interpretation of such ID. The newly developed method paves the way to push the investigation even further, toward the extension to discrete metrics of distance-based algorithms and routines that are, nowadays, consolidated in the continuum, such as density estimation methods, or clustering algorithms.

The supporting data for this Letter are openly available from [33].

The authors thank Antonietta Mira, Alex Rodriguez, and Marcello Dalmonte for the fruitful discussions. A. G. and A. L. acknowledge support from the European Union's Horizon 2020 research and innovation program (Grant No. 824143, MaX "Materials design at the eXascale" Centre of Excellence). The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of Banca d'Italia.

I. M., A. G., and A. L. designed and performed the research. All authors wrote the paper. J. G. designed the application on genomics sequences.

[*] laio@sissa.it

[1] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, Artif. Intell. Rev. **53**, 907 (2020).

[2] A. Jović, K. Brkić, and N. Bogunović, *A Review of Feature Selection Methods with Applications* (IEEE, New York, 2015), pp. 1200–1205, 10.1109/MIPRO.2015.7160458.

[3] Y. Bengio, A. Courville, and P. Vincent, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).

[4] T. Mendes-Santos, X. Turkeshi, M. Dalmonte, and A. Rodriguez, Phys. Rev. X **11**, 011040 (2021).

[5] T. Mendes-Santos, A. Angelone, A. Rodriguez, R. Fazio, and M. Dalmonte, PRX Quantum **2**, 030332 (2021).

[6] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, Chem. Rev. **121**, 9722 (2021).

[7] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan, Adv. Neural Inf. Process. Syst. **32**, 6111 (2019).

[8] D. Doimo, A. Glielmo, A. Ansuini, and A. Laio, Adv. Neural Inf. Process. Syst. **33**, 7526 (2020).

[9] S. Recanatesi, M. Farrell, M. Advani, T. Moore, G. Lajoie, and E. Shea-Brown, arXiv:1906.00443.

[10] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza, Math. Probl. Eng. **2015**, 759567 (2015).

[11] F. Camastra and A. Staiano, Inf. Sci. **328**, 26 (2016).

[12] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications* (John Wiley & Sons, New York, 2004).

[13] A. Block, W. von Bloh, and H. J. Schellnhuber, Phys. Rev. A **42**, 1869 (1990).

[14] P. Grassberger, Int. J. Mod. Phys. C **04**, 515 (1993).

[15] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).

[16] K. Christensen and N. R. Moloney, *Complexity and Criticality* (World Scientific Publishing Company, Singapore, 2005), Vol. 1.

[17] L. Niemeyer, L. Pietronero, and H. J. Wiesmann, Phys. Rev. Lett. **52**, 1033 (1984).

[18] A. Kosior and K. Sacha, Phys. Rev. B **95**, 104206 (2017).

[19] E. Facco, M. D'Errico, A. Rodriguez, and A. Laio, Sci. Rep. **7**, 1 (2017).

[20] J. Theiler, J. Opt. Soc. Am. A **7**, 1055 (1990).

[21] M. Möller, W. Lange, F. Mitschke, N. Abraham, and U. Hübner, Phys. Lett. A **138**, 176 (1989).

[22] E. Ehrhart, *International Series of Numerical Mathematics* (Birkauser Verlag, Basel-Stuttgart, 1977), Vol. 35.

[23] D. Moltchanov, Ad Hoc Networks **10**, 1146 (2012).

[24] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.130.067401 for detailed mathematical derivations and further examples.

[25] Eugène ehrhart—publications 1947–1996, http://icps.u-strasbg.fr/ clauss/Ehrhart_pub.html.

[26] M. Beck and S. Robins, Choice Rev. Online **45**, 45 (2007).

[27] Qiita Spots Patterns - Sponge and Algal Associated Water Microbiome, https://qiita.ucsd.edu/study/description/13596.

[28] E. Bolyen *et al.*, Nat. Biotechnol. **37**, 852 (2019).

[29] M. W. Gray, D. Sankoff, and R. J. Cedergren, Nucl. Acids Res. **12**, 5837 (1984).

[30] C. R. Woese, O. Kandler, and M. L. Wheelis, Proc. Natl. Acad. Sci. U.S.A. **87**, 4576 (1990).

[31] W. G. Weisburg, S. M. Barns, D. A. Pelletier, and D. J. Lane, J. Bacteriol. **173**, 697 (1991).

[32] J. Jovel, J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A. L. Mason, K. L. Madsen *et al.*, Front. Microbiol. **7**, 459 (2016).

[33] A. Glielmo, I. Macocco, D. Doimo, M. Carli, C. Zeni, R. Wild, M. d'Errico, A. Rodriguez, and A. Laio, Patterns **3**, 100589 (2022).