scientific reports

Check for updates

OPEN

Intrinsic dimension as a multi-scale summary statistics in network modeling

Iuri Macocco[1], Antonietta Mira[2,3] & Alessandro Laio[1,4]✉

Complex networks are powerful mathematical tools for modelling and understanding the behaviour of highly interconnected systems. However, existing methods for analyzing these networks focus on local properties (e.g. degree distribution, clustering coefficient) or global properties (e.g. diameter, modularity) and fail to characterize the network structure across multiple scales. In this paper, we introduce a rigorous method for calculating the intrinsic dimension of unweighted networks. The intrinsic dimension is a feature that describes the network structure at all scales, from local to global. We propose using this measure as a summary statistic within an Approximate Bayesian Computation framework to infer the parameters of flexible and multi-purpose mechanistic models that generate complex networks. Furthermore, we present a new mechanistic model that can reproduce the intrinsic dimension of networks with large diameters, a task that has been challenging for existing models.

Network theory is a popular and powerful tool for modeling and explaining the emergent properties of various complex systems in many different fields, such as economics, social sciences, biology, transportation, neuroscience, and beyond[1–4]. Considerable efforts have been devoted to designing models capable of grasping and reproducing the properties of real networks as closely as possible. For example, a mechanistic approach can be followed, in which one specifies a set of microscopic rules (usually involving the creation/deletion of nodes and links) used to grow or evolve the network[5,6]. It is—typically—relatively simple to produce samples of networks from a particular mechanistic model. Although such a framework allows the incorporation of specific knowledge domain and expertise to design network formation rules, mechanistic models are still burdened by the human understanding of the problem, meaning that while generative routines can be very rich and flexible, they are inherently limited and capable of synthesizing only some particular ensembles of networks[7,8]. A more general alternative is offered by probabilistic approaches, such as the so-called Exponential Random Graph Model[9], where networks are generated according to a given probability measure. Network generation is tightly entangled with the related topic of network comparison. Indeed, only by introducing a specific similarity measure it is possible to say whether the artificial network generated by a model is akin to an observed real-world network. Many different options are available to compare two networks, from measuring the distance between adjacency matrices to more sophisticated metrics and observables, often based on information theory principles[10–14]. Some of the most common observables generally focus on local (e.g. degree distribution, local clustering) or global (e.g. diameter—the largest shortest path—) properties. Mechanistic generators try to replicate a subset of network properties but, unfortunately, a proper sufficient subset of observables that would ensure a realistic correspondence of the synthesized networks with the observed ones is still lacking. In this work, we propose using the intrinsic dimension (ID) of the network at different scales to compare simulated and real world networks. The ID of a dataset was originally introduced for characterizing dynamic systems, but has now become a key concept in the realm of manifold learning. In fact, data points are typically embedded in high-dimensional spaces characterized by a large number of features, among which a certain degree of (possibly nonlinear) interdependence exists. For this reason, the data points are believed to belong to a lower dimensional manifold defined by a small number of independent variables. The number of such variables is the ID. Indeed, in most real-world datasets it is possible to describe the data, at least locally, using a number of coordinates which is much lower than the number of features[15–17]. Very interestingly, the estimated ID depends on the scale at which a given system or dataset is observed, reflecting the complexity of the data structure at different resolution levels. For example, data laying on a line and perturbed by d-dimensional noise have an ID of order d at a small scale, and of order one at a large

[1]International School for Advanced Studies (SISSA), Via Bonomea 265, 34136 Trieste, Italy. [2]Faculty of Economics, Euler Institute, Università della Svizzera italiana, Via Buffi 13, 6900 Lugano, Switzerland. [3]Department of Science and High Technology, Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy. [4]The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34014 Trieste, Italy. ✉email: laio@sissa.it

scale. The motivating idea of this work is to exploit the rich information encoded in this scale dependence to characterize the properties of unweighted networks.

The concept of estimating the dimensionality of a network is not new; the first attempts to characterize the dimension of a network can be found in[18,19] and refined in successive studies[20,21]. This concept is partially related to fractality and self-similarity, properties that have been extensively explored in networks[22,23] and that can currently be characterized by various mathematical tools[24–26] often relying on instruments inspired by manifold learning techniques such as the Box-Counting[27,28] or the Correlation Dimension[29,30]. However, to our knowledge, in all methods introduced so far the discreteness of the distances between the nodes was not taken explicitly into account. For example, renormalization methods for determining the fractal dimension simply cover the nodes without considering whether the edges are weighted or not, often assuming that the distance between nodes is a real number. If applied to unweighted networks, where the distance between vertices are computed as the shortest path and the edges are unweighted, these approaches may be affected by systematic errors. Moreover, and possibly more importantly, available approaches do not allow estimating the ID as an explicit function of the scale, a feature that is essential if one wants to use the ID as a network fingerprint. To overcome such limitation, we propose the use of an ID estimator specifically built for data spaces in which the distances can only take discrete values[31]. Some examples of the ID as a function of the scale, which we name ID signature, estimated by our approach are reported in Fig. 1. The networks were obtained by first sampling points uniformly at random on a low-dimensional manifold which is then embedded in a high-dimensional space by adding coordinates that are all zero and, subsequently, all coordinates are perturbed with Gaussian noise. Finally, the network is created by connecting each point to a number of neighbors that is fixed, on average, to a given value. In the blue and green networks, only the first coordinate is uniformly sampled over the interval (0,10) and 99 fictitious zero coordinates are added; the added Gaussian noise has standard deviations of, respectively, $\varepsilon = 10^{-4}$ and $\varepsilon = 5 \cdot 10^{-3}$. The ID at a large scale is close to one, corresponding to the dimensionality of the underlying manifold emerging when noise becomes irrelevant; at a smaller scale, the ID is determined by the number of neighbors that are linked to create the network structure, namely, by the degree, and by the variance of the Gaussian noise. The green network, in which the average degree is fixed to 10 and the variance of the noise is large, the ID is 5 at distance one, but then it becomes significantly larger, and reaches the value of 1 only at a scale of ~ 10. For the blue network, which has the same average degree as the green one but a smaller noise, the ID is also equal to 5 at distance 1 but then quickly plateaus to 1. Finally, the orange network, for which both the first and second coordinates are uniformly sampled in the interval (0,10), has the average degree fixed at 6 and a small noise, so that the ID is ~ 3 at a small distance, and very quickly converges to two.

The rich behavior of the ID as a function of the scale observed in artificial networks that emerges from Fig. 1 prompted us to introduce an approach that allows inferring the parameters of a mechanistic model able to reproduce a target ID curve. This is achieved by wrapping a generative model within the framework of the Approximate Bayesian Computation[32,33] (ABC). We will show that networks generated using ABC in such a way that their ID signature reproduces that of a target network will be statistically similar/close to the target network

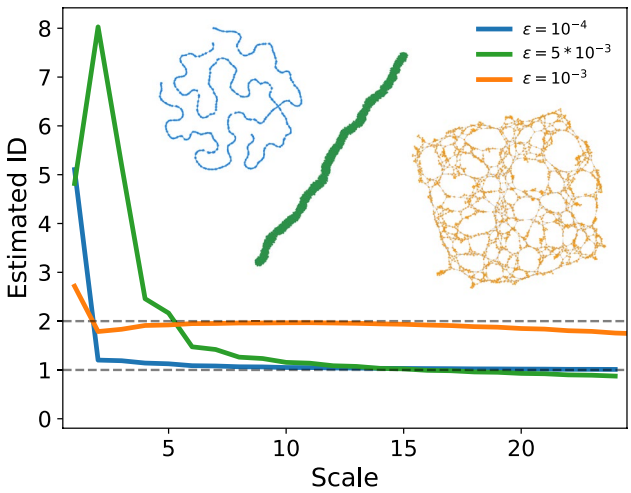

Figure 1. The ID signature of networks built upon points embedded in a metric space is consistent with the ID of the underlying manifold from which the points were extracted. Points are uniformly sampled in the interval (0,10) (blue and green curves) or in a square with a side length of 10 (orange curve), embedded in 100 dimensions by adding 99 or 98 zeros, and then Gaussian noise with standard deviation ε is added to all the coordinates. We then create the network by connecting each 100-dimensional point to a random number of neighbors extracted from a Poisson distribution with given intensity (i.e. mean value) λ. λ is 10 for the green and blue networks and 6 for the orange one. While at high scales the ID signature reaches a plateau corresponding to the expected value (namely 1 or 2), at small distances the ID depends on the given combination of ε and λ. In particular, the larger is the noise, the greater the dimensionality estimated at small scales, as the low-dimensional structure becomes temporarily hidden from the nonnegligible noisy coordinates.

according to degree distribution, closeness, betweenness, clustering coefficient and page rank, all observables that are not directly controlled for in ABC. We apply our algorithm on benchmark systems, comparing its performance on simple generative models, and show that ABC using the ID signature as summary statistics, allows the estimation of the ground truth generative parameters. This occurs even for more sophisticated and flexible mechanistic models, with up to 8 parameters. For some classes of networks, we find that the ID signature alone, even if well reproduced, does not allow us to uniquely identify the original parameters with a sufficiently narrow posterior; it leads, instead, to a class of estimates—a manifold in the space of parameters—that, nonetheless, always includes the generative parameters associated with the observed network. We also find that, for other networks characterized by a large diameter, it is impossible to reproduce the reference ID signature with generative models available in the literature and that we tested. Therefore, we developed a new generative network algorithm in which the ID at different scales is dynamically controlled while the network is built. We will show that this approach allows the generation of ensembles of networks whose properties are statistically similar to real-world networks, even when standard generative models fail.

Methods

In this section, we present the three ingredients of our approach: the discrete intrinsic dimension estimator, the ABC algorithm and the ID-guided generative network model.

Estimating the ID of a graph

We infer the ID of a network from the probability distribution of the distances between the nodes as measured by the number of links to go from one node to another following the shortest path. Since these distances, in an unweighted network, can only take discrete values, we use I3D[31], an approach specifically developed for estimating the ID in datasets in which the features can only assume integer values. In this approach, one chooses two integers R_1 and $R_2 > R_1$, and counts, for each data point i, the number of data points whose L^1 distance from i is smaller than or equal to R_2 (denoted by k_i), and the number of data points whose distance from i is smaller than or equal to R_1 (denoted by n_i). In Ref.[31,34] it is proven that, under some regularity assumptions, n_i is a realization of a binomial random variable where k_i is the number of trials, and the success probability p depends only on R_1, R_2, and the dimension d of the square lattice which contains the data: $n_i|k_i \sim \mathrm{Binomial}(\mathrm{k_i}, \mathrm{p(d, R_1, R_2)})$. An explicit formula for p can be derived by the Ehrhart theory of polynomials[35,36]. In particular, $p(d, R_1, R_2) = V_\diamond(R_1, d)/V_\diamond(R_2, d)$ where $V_\diamond(R, d)$ is the number of points that would be observed within a distance R on a square lattice of dimension d:

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

where ${}_2F_1(\cdot, \cdot, \cdot, \cdot)$ is the ordinary hypergeometric function. Extending the reasoning from a single reference point i to the whole dataset, and assuming independence among the N random variables n_i, one can write the conditional probability of observing the vector of n_i-s given the values of k_i-s and given p:

$$\mathcal{L} = \prod_{i=1}^{N} \mathrm{Binomial}(\mathrm{n_i}|\mathrm{k_i}, \mathrm{p(d, R_1, R_2)}). \tag{2}$$

where $\mathrm{Binomial(n|k, p)}$ denotes the probability mass function of a $\mathrm{Binomial(k, p)}$ random variable evaluated in n. One then estimates the ID by maximising $\mathcal{L}$ with respect to d (or, alternatively, by taking the average of the posterior, which, assuming a Beta conjugate prior, is a Beta distribution[31]). This amounts to finding the root of

$$\frac{V_\diamond(R_1, d)}{V_\diamond(R_2, d)} - \frac{\langle n \rangle}{\langle k \rangle} = 0. \tag{3}$$

where $\langle n \rangle$ e $\langle k \rangle$ are averages computed over all nodes in the network. By varying R_2 and imposing $R_1 = r\, R_2$, where $0 < r < 1$ is the only free parameter of the model, one obtains the value of the ID at different scales R_2. In the rest of the work, we set $r = 0.5$.

The approach we just described allows estimating the ID using only distances, and can therefore be applied with no modification to estimate the *ID* of an unweighted network. The ID estimated in this manner can practically be thought of as the dimensionality of the lattice on which a subgraph centered on node i and including only the nodes up to a distance R_2 could be (approximately) embedded. While this ID can in principle be different in different nodes, we here assume it to be constant. The goodness of the estimate is verified, for each R_2, by model validation (see Supp. Inf.). In Supp. Inf. we also show that correlation effects, related to the values of n_i-s on different nodes of the same network, do not significantly affect the estimate of the *ID* obtained by maximizing the likelihood (2), which is written under the assumption of independence of the different counts.

Importantly, as shown in the examples in Fig. 2, the dimensionality of such an embedding lattice can be different at different scales. For instance, in the network representing yeast protein interactions, the ID at a small scale is of order 4, while, at a larger scale, decreases rapidly to 0. In fact, at large scales, every system looks like a zero-dimensional point. In general, each network has a peculiar and specific ID signature, which, as we will see, can be considered as a fingerprint that allows identifying many properties of the network itself. Differently from other observables that typically focus on local (e.g. degree distribution, local clustering) or global (e.g. diameter) properties, the ID signature instead spans short-, meso- and long-range scales.

When dealing with geometrical networks, namely, graphs built from points contained in a Riemannian manifold of a given dimension, the ID has a proper quantitative meaning. As the scale becomes larger, the ID tends to

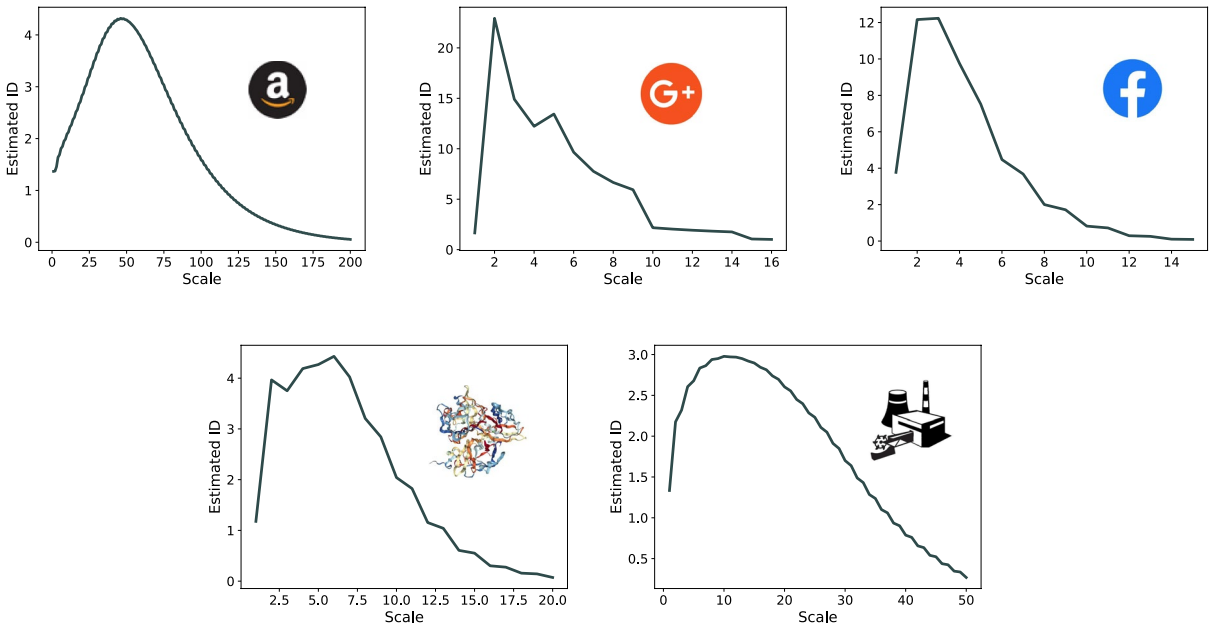

Figure 2. ID signatures for real-world networks. Such a variety of ID profiles makes this observable a good candidate for characterizing different graphs. From the top left, clockwise direction: Amazon recommendations, Google+, Facebook companies, yeast proteins, US power grid. All networks were downloaded from[49] and some of their summary statistics can be found in Table 1.

the dimension of the underlying manifold, as shown by the plateaus in Fig. 1. These plateaus identify an ID that is constant over a wide range of scales. However, one of the strengths of network theory is its ability to describe processes and systems that cannot be thought of as embedded in an underlying geometrical manifold, defined by explicit coordinates. In these cases, the ID curve is typically a complex function of the scale, exhibiting no clear plateau. As will be discussed in further detail, the ID curve is extremely sensitive even to tiny changes in the network topology. For example, it is easy to change it almost entirely by adding even a single extra link. The reason behind this sensitivity is the complex relationship between the edge structure and the distances between nodes: Moving a single edge can create a link between two clusters of nodes that would otherwise be far apart, dramatically changing the vertex distance matrix, even if, for instance, the degree distribution is kept identical.

Our approach is based on the shortest path distances between nodes, meaning that self-loops or parallel edges (i.e., more edges connecting a couple of vertices) would not change the ID estimation. The same algorithm can be applied to directed networks, as the only difference from undirected ones is that the distance matrix is not symmetric. However, to keep simple and intuitive examples, we focus on simple (loop-less and single-edge) undirected networks.

Approximate Bayes computation

To illustrate the usefulness of the ID signature as a fingerprint of a network, we demonstrate that it can be used as a summary statistics to estimate the parameters of a generative model capable of producing networks with properties that are very similar to those of a given target network. Among the available methods to estimate the parameters of the generative processes, we resorted to Sequential Monte Carlo Approximate Bayes Computation[37,38], a flexible and adaptable scheme that, with respect to other estimation methods, also provides uncertainty quantification. Approximate Bayes Computation (ABC) has been successfully employed in many different fields[32,33] and in the context of network theory[39–41]. The simplest ABC algorithm is the so-called Rejection ABC, which aims at inferring a posterior distribution $f(\theta|x_0)$ for a set of parameters θ given some observed data x_0 when the likelihood of the model cannot be evaluated explicitly but can be easily sampled. Given a putative vector θ^* extracted from a prior distribution $\pi(\theta)$, one generates a simulated network x from the model (encoded by a likelihood function f) conditioned on the sampled value of the parameter, $f(x|\theta^*)$. The proposed θ^* is accepted if the simulated data are "close" to the observed data. Such closeness can be assessed by computing the distance between some summary statistics $S(x)$. One thus approximates the posterior according to $f(\theta|x_0) \sim f(\theta \mid \rho(S(x), S(x_0)) < \varepsilon)$ where ρ is a suitable distance measure and ε is a tunable tolerance. In the limit of $\varepsilon \to 0$ and if $S(\cdot)$ is a sufficient summary statistics the approximate posterior becomes exact. This procedure alone is however very inefficient, and many improvements have been proposed[42]. As already mentioned, here we will exploit the Sequential Monte Carlo extension of ABC (see[43] for details about its advantages and drawbacks). Briefly, within this framework, a population of parameters $\theta_1, \ldots, \theta_n$ (termed particles) evolves from an initial prior distribution through a sequence of intermediate distributions that tend to converge to the true posterior. Along this history, the tolerance ε is reduced at each step according to a tunable scheme, resulting in Algorithm 1:

Algorithm 1 The Sequential Monte Carlo ABC procedure

Input: Number of nodes N, x_0 observed network, reference summary statistics $S(x_0) = \mathrm{ID}_R(x_0)$, Prior distribution $\pi(\theta)$, target tolerance threshold ε, number of particles n
Output: Approximate posterior samples $\{\theta^{(t,n)}\}_{t=1}^{T}$

$g_0(\theta) = \pi(\theta)$
Initialize ε_0 via calibration samples,
$t \leftarrow 1$
while $\varepsilon_t > \varepsilon$ **do**
    $n_{count} \leftarrow 0$
    $\rho_t(x_0) = [\,]$
    while $n_{count} < n$ **do**
        Sample $\theta^{\mathrm{prop}} \sim g_t(\theta|\theta^{(t-1,n)})$ (g_t is the proposal distribution for parameters)
        Generate network $x \sim f(x|\theta^{\mathrm{prop}})$
        Compute distance $\rho(S(x), S(x_0))$ between the summary statistics.
        if $\rho(S(x), S(x_0)) \leq \varepsilon_t$ **then**
            Accept θ^{prop}
            $n_{count} \leftarrow n_{count} + 1$
            Add $\rho(S(x), S(x_0))$ to $\rho_t(x_0)$
        end if
    end while
    $\varepsilon_{t+1} \leftarrow \mathrm{median}(\rho_t(x_0))$
    $t \leftarrow t + 1$
end while
return $\{\theta^{(t,n)}\}_{t=1}^{T}$

where $\rho_t(x_0)$ are the distances of the summary statistics of the accepted simulated networks from the reference summary statistics $S(x_0)$ at generation t. The algorithm runs until $\varepsilon_t > \varepsilon$, the "calibration sample" consists of accepting n particles extracted from the prior, and ε_1 is obtained from this sample as the median$(\rho_0(x_0))$. The proposals θ^{prop} at time t are extracted by means of a kernel density estimate, so that $g_t(\theta|\theta_{t-1})$ is a multivariate Gaussian distribution with mean and covariance given by the mean and covariance of the particles accepted at step $t-1$ (further technical details are provided in the SI).

In principle, instead of resorting to a Bayesian approach, the estimation of the parameters of a generative model can be tackled by means of optimizations methods, namely the Stochastic Path Integral[44] or the Pareto Simulated Annealing[45]. Identifying the most suitable one for the specific task of matching the ID signature of a generative model will the object of future research. For the time being we decided to stick to the SMC-ABC as, firstly, it provides statistically rigorous uncertainty quantification, a feature which is often absent in optimization methods. Secondly, it allowed us to reliably explore all the relevant parameter space using moderate computational resources through a simple and straightforward integration of our model within the well documented `pyABC` library[46].

ID-guided generative model

Our main idea consists of using the ID signature as a vectorial summary statistics for the ABC procedure, and to use the L^∞ metric between the reference (observed or ground truth) ID signature and the signature associated with the generated networks x:

$$\rho(S(x), S(x_0)) = \max_{R<\Delta(x)} |\mathrm{ID}_R(x) - \mathrm{ID}_R(x_0)| := \mathcal{D}(x, x_0) \tag{4}$$

where $\Delta(x)$ is the diameter of the network and $\mathrm{ID}_R(x)$ is the ID of network x computed at $R_2 = R$.

As shown in the "Results" section, the classical ABC procedure, with the ID as a summary statistics, works very well in many test cases, in artificial settings when the target network is obtained through a generative algorithm, but also for real-world networks characterized by a low diameter. However, generative models available in the literature that we have checked all fail to reproduce the ID signatures of real-world networks *with low IDs and high diameters*.

Name	N	E	l	Δ	D
Amazon recom.	91,814	125,704	51.2	200	5
Google+	23,629	39,194	4.1	8	2761
Fb companies	14,113	52,126	5.3	15	215
Yeast proteins	2115	2203	7.1	19	56
US power grid	4941	6594	18.9	46	19

Table 1. Comparison of networks in Fig. 2 according to some characteristics. N: number of nodes, E: number of edges, l: average shortest path, Δ: diameter (of the connected component), D: largest degree.

In some cases, the standard generative processes fail, possibly due to a highly nonlinear relationships between the links (see Results for specific examples). To address this problem, we designed a new network generative process that, during the construction, compares the current ID signature with the reference signature. In particular, for each move proposing the addition of an edge, we compute the ID signature, and the move is accepted following the Metropolis-Hastings algorithm, i.e., with a probability proportional to:

$$\min\left(1, \exp\left[\beta\left(\mathcal{D}(x', x_0) - \mathcal{D}(x, x_0)\right)\right]\right) \tag{5}$$

where β plays the role of the inverse of a temperature, x_0 is the reference network, x is the last accepted network and x' is the network with the newly proposed edge.

This procedure is similar to that proposed for the dk-models in[47]. In that paper, the authors, in order to generate a random graph that tries to reproduce an observed one, take into account the original average degree, the degree distribution, the joint degree distribution, the average clustering coefficient and the degree-wise average clustering coefficient, and perform edge swapping while targeting the aforementioned statistics. They show that, in several cases, by limiting the dk-series at the 2.5k level (which means taking into account local properties) also other network observables at the meso- and macroscopic scales are reasonably well matched. However, already in the supplementary information of the same paper, it is reported that this does not happen for the BRAIN network, in which betweenness and shortest path length are not reproduced. The same occurs for the US power grid network[48]—one of our test cases—even if the 3k statistics is also accounted for. We think this is not by chance. Indeed, such graphs are characterized by a particularly high diameter, a feature that, as we will show, cannot typically be obtained by available models.

In this work, we focus exclusively on the ID signature as our unique summary statistics, and let the network grow from scratch. In each step, the network growth process is guided by meaningful and interpretable edge addition actions. Also this model is integrated within a Sequential Monte Carlo Approximate Bayesian Computation framework to estimate the optimal β and the vector of probabilities associated with the edge addition actions.

Results

ID signature of real-world networks

The first result we present concerns the staggering variety of ID signatures of real-world networks. In Fig. 2, we report five representative examples of social networks, a protein interaction network and infrastructure networks. The ID at distance 1 is directly related to the average degree. Indeed, let us consider Eq. (3): for $R_1 = 0,\ R_2 = 1$ we get $\frac{1}{2d+1} - \frac{1}{\langle k_1\rangle+1} = 0$, from which $d = \langle k_1\rangle/2$, where k_R indicates the distribution of neighbors at distance R, so that $P(k_1)$ is the canonical degree distribution and $\langle k_1\rangle$ is thus the average degree. For larger R the estimator considers the average number of neighbours up to that distance. The common (more or less pronounced) observed increasing trend of the ID at a small scale is due to the exponential growth of the number of neighbors of each vertex with the radius R. In particular, a sharp rise to high ID values in this regime is typically associated with the presence of hubs, nodes characterized by a high degree that foster high connectivity and enlarge the size of neighborhoods at small distances. Accordingly, for vertices connected to such hubs, the number of neighbors at distance 2 or 3 is exponentially larger than their degree. This is the case for Google+ and Facebook companies networks, reported in Fig. 2, where the degrees are (approximately) power-law distributed and the IDs show a narrow peak, reaching values of 20 and 12 respectively. Conversely, in cases where the degree distribution is more uniform and hubs are absent, the ID signature at the mesoscale region is smoother and can present a quasi plateau across a wide range of distances. This occurs, for instance, in the yeast protein network and in the US power grid network, where the ID is, respectively, of order 4 for distances between 2 and 7 and of order 3 for distances between 7 and 17. At scales comparable with the average path length, the ID curve typically peaks and then starts declining, since the growth of neighbourhoods' size becomes subpolynomial. The largest scale that is still meaningful coincides with the largest shortest path, which is typically called diameter.

The networks were downloaded from[49], where further references and details can also be found.

ID-guided ABC for the Erdös–Rényi generative model

To assess the goodness of the ID signature as a summary statistics, we first check whether it can be used to retrieve the parameters chosen to create a reference network via a simple model. All the simulations involving ABC were performed using the `PyABC` package[46]. We first consider the Erdös–Rényi[50] (ER) model, one of the simplest and most common random-graph models. According to the ER model, each possible edge of a network with N

vertices is independently added with probability p, implying an average number of edges $\langle E \rangle = \binom{N}{2} p$. For our first experiment we set $N = 300$, $p = 0.01$ (and thus $\langle E \rangle = 448.5$) and extract the ID signature of a single ER graph realization. This is the reference summary statistics that we want to reproduce. In this first test we simply check whether it allows us to properly estimate the value of p used to generate the related network.

The results are reported in the first row of Fig. 3. The panel on the left shows the evolution of the average ID (with its standard deviation) throughout successive ABC generations; the central panel reports the successive posterior approximations of p; the panel on the right displays how convergence to the target ε occurs in 10 generations, following exponential decay. In particular, we observe that the average ID of the sampled networks is far from the target for the first 3 generations, and, accordingly, the associated posterior distribution is still broad. Together with the decreasing of the distance between the reference ID signature and the one of the sampled networks, we can see a sharpening of the posterior around the ground truth value $p = 0.01$. For readability reasons, we reported the evolution of the average ID and the posterior up to $t = 5$, as successive generations are practically indistinguishable in the plots, and would make the figure less clear.

ID-guided ABC for artificial networks

The results obtained for the ER model were expected, as the average number of links $\langle E \rangle$ or, equivalently, the average degree, which are sufficient statistics for this model, are encoded in the ID signature at distance $R = 1$. Next, we considered less trivial generative models: the Non-Linear Preferential Attachment[52] (NLPA), the Watts–Strogatz[53] (WS), and the planted partition[54] (PP). In such models, trivial sufficient summary statistics are not known. Here, we show that ABC, with the ID signature as the unique summary statistics, returns samples from the posterior that are centered around the parameters chosen as the ground truth. In the second row of Fig. 3 we report the results obtained for a PP graph, a model meant to build interacting communities. Once the number of communities and the number of elements per community are fixed, two more parameters, p_{in} and p_{out}, regulate the probability of connecting vertices within a community and among communities. We followed the same Algorithm 1, by appropriately considering the higher dimensionality of the parameter space (equal to 2 in this case). The convergence of ε_t to the target ε is exponential, similarly to the ER case and, thus, not reported. Differently from the previous example, we can appreciate how the average ID signature is apparently close to the reference one, already from the first generation (left panel). However, the wide standard deviation implies a diversified population of graphs and, accordingly, a broadened posterior distribution. The narrowing of the standard deviation is then paired with the concentration of the posteriors around the ground truth parameters. The fact that, to achieve reasonable precision on the posterior, one needs the whole population of sampled graphs to have similar ID signatures hints at the meaningfulness and power of such summary statistics, which is sensitive even to small variations of the generative parameters. The results for the other mentioned models are qualitatively similar and are thus reported in the SI.

To more closely mimic the generative mechanism behind real-world networks, we then considered more flexible and rich generative models. Many methods have been proposed throughout the years, including the dk-random graphs[47], Chung-Lu[55,56] and the Exponential Random Graph Models (ERGM)[9]. We here use the so-called Action Based Network Generators (ABNG)[51], a mechanistic model that proposes the addition of edges according to intuitive and interpretable actions, which are based on well-known network properties. At each iteration of the generative process, a vertex is randomly chosen and a new link toward another node is added according to the probability associated with different possible actions. For instance, one of those consists of creating a triadic closure; another builds the edge in relationship to the degree of the target or to the degree of the target's neighbors. For different combinations of the aforementioned probabilities associated with the actions, the model gives rise to networks with radically different structures. In Ref.[51] the Authors show the flexibility of this model by reproducing the most common random graphs and a wide class of real-world networks. In particular, they estimate the probabilities of actions through Pareto Simulated Annealing[45], fitting the following summary statistics: degree distribution, page rank, betweenness centrality and clustering coefficient. In contrast, as already stated, we calibrate this generative model using ABC with the ID signature as the only summary statistics.

As a preliminary step, we verify that, given certain ground truth probabilities and a reference network realization, we could reproduce its ID and correctly estimate the original parameters. In the bottom row of Fig. 3, we report the results obtained on artificial networks generated using 6 different actions. As in previous cases, the reference ID is perfectly retrieved and the ground truth parameters are within the confidence level of the posteriors.

In this case, it must be noted that we are trying to infer 6 parameters, meaning that at least 6 constraints need to be enforced. For this reason, it is important that the scale at which the ID can be computed is larger than the number of parameters one wants to infer. Actually, since the IDs computed using Eq.(3) are exploiting cumulative neighborhoods, it means that the information extracted at a given distance r will also be used in the estimations for $R > r$. As a consequence, in principle, one wants the diameter of the network to exceed (possibly by far) the number of parameters to be estimated.

ID-guided ABC fails to fit the ID of high-diameter networks

As a final step, we then moved to real-world networks, where a ground truth generative model is not available. We start from a subset of the Facebook network, obtained by selecting a random vertex and its 500 nearest-neighbors, which are linked if the edge is also present in the original graph. The ID of the subset network resembles the one computed on the whole-network: a sharp increase in the ID at small scales and a small diameter, of order ~ 10. The results of the ABC protocol are shown in the first row of Fig. 4 (for clarity and readability, we present only the last generation of the ABC routine). One can appreciate how the ID (in blue) is fairly reproduced but does not reach the target $\varepsilon = 0.05$. To assess the quality of the ensemble of networks produced in the last generation

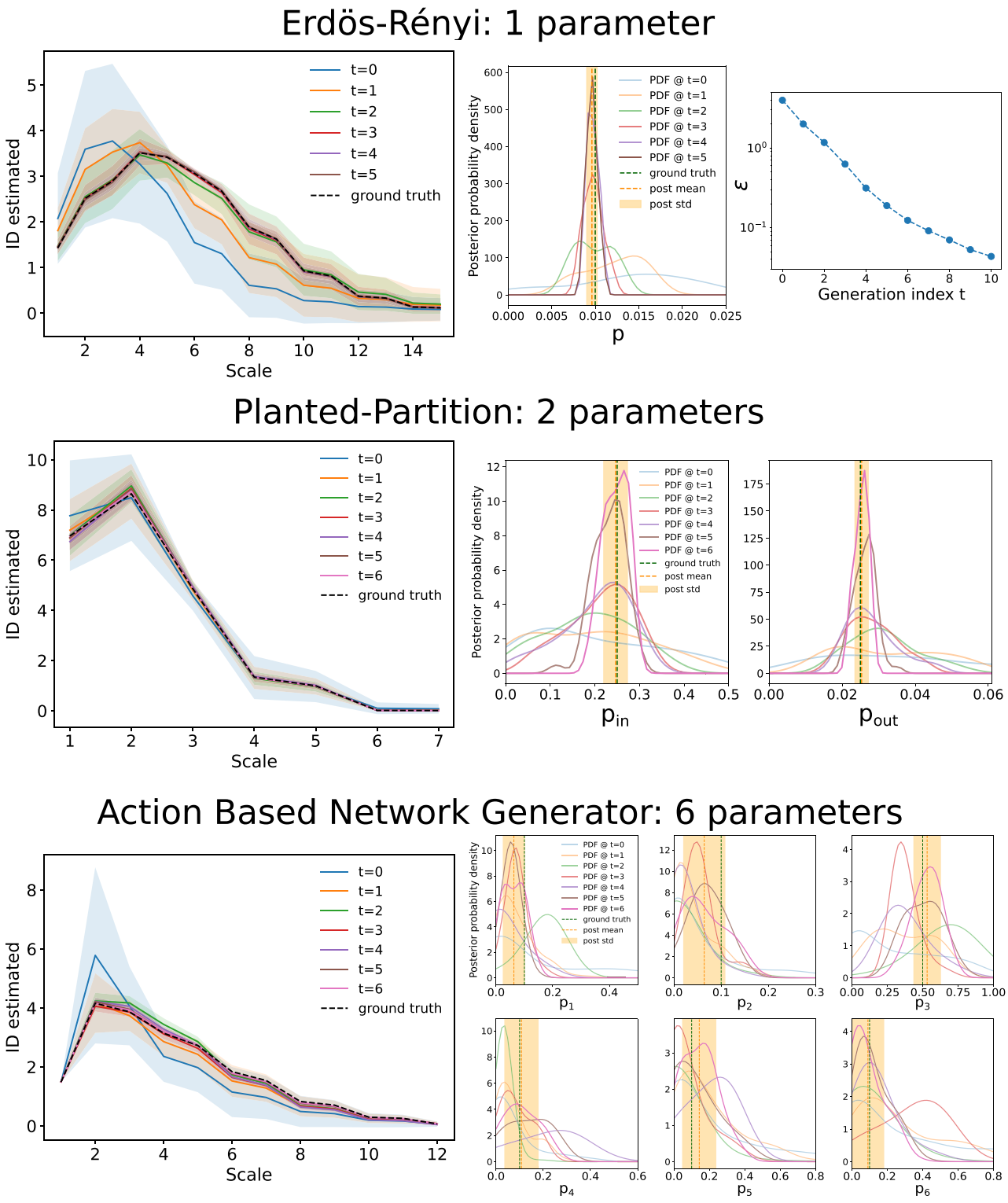

Figure 3. The ID signature used as a summary statistics allows us to retrieve the ground truth parameters used to generate the reference network for different generative mechanistic models. In each row, the left panels show the evolution of the ID through the SMC generations, while on the right we display the successive posteriors associated with the accepted particles. The evolution of ε_t is reported only in the first row, as the same behaviour is observed in all other cases. For all models, the target tolerance threshold is $\varepsilon = 0.05$, and the number of particles is $n = 50$. The first row shows the Erdös–Rényi model (depending on 1 parameter p), with the following simulation details (with reference to Algorithm 1): number of nodes $N = 300$, $p = 0.01$, prior distribution $\pi(p) = \text{unif}[0, 0.025]$. The second row displays the Planted Partition model (2 parameters, see text for model details), where we set the number of communities $l = 10$, the number of nodes per community $k = 30$, $p_{in} = 0.25$ and $p_{out} = 0.025$. The third row is a realization of the ABNG[51] model with 6 parameters. The prior is given by a uniform distribution on the simplex defined by $\sum_i^6 p_i = 1$.

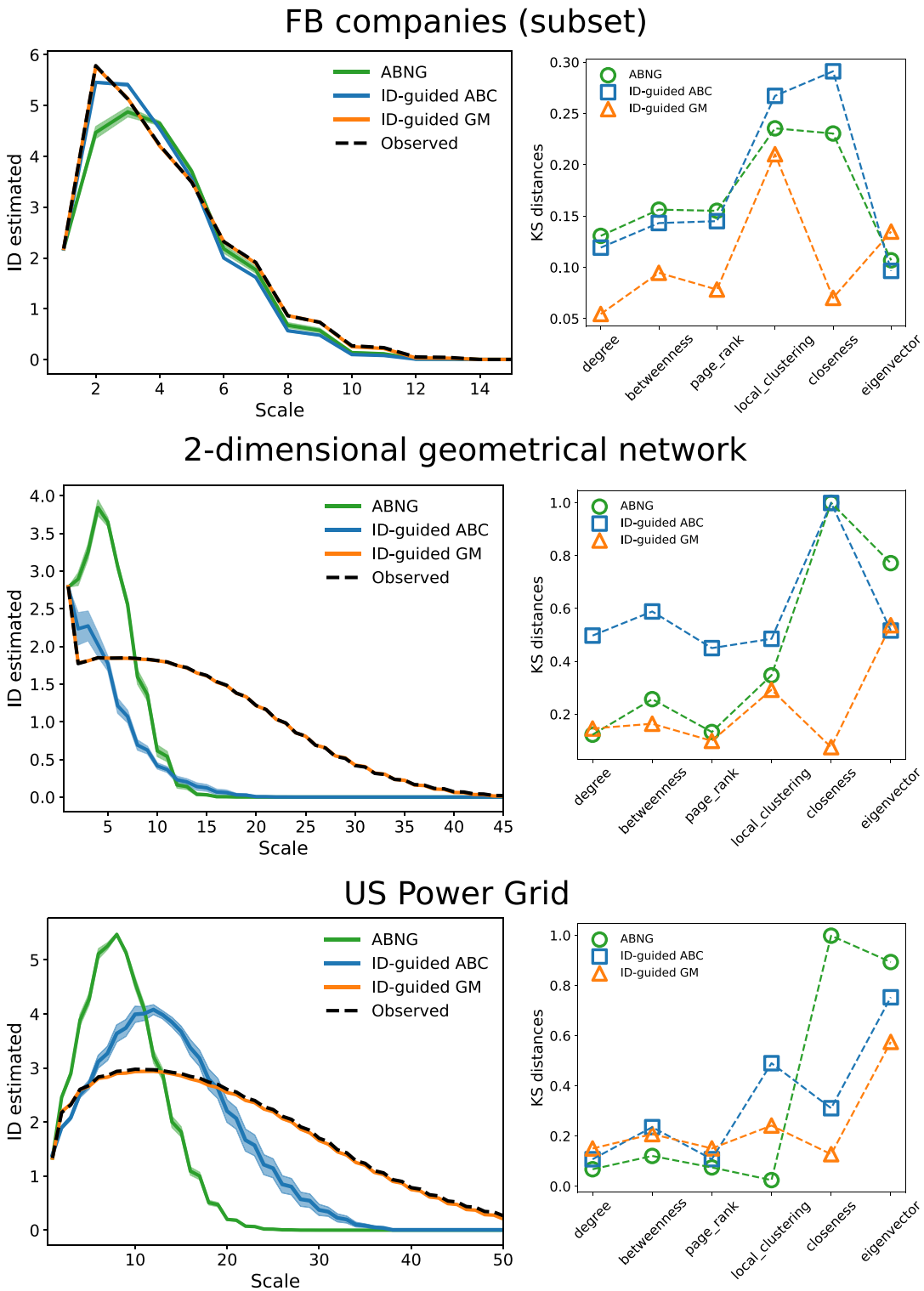

Figure 4. Performances of ABNG against ID-guided ABC and ID-guided GM for two real-world and one geometric networks. The panels in the left column compare the different ID signatures, while the panels on the right represents the average of the KS distances between the ensemble of sampled networks against the reference networks for 6 different typical graphs properties. The first row concerns a subset of the Facebook company network. In this case, the ABNG model allows to reproduce quite faithfully the ID even if fitted onto observables different from the ID. The ID-guided ABC gets closer to the ID but does not reach the target $\varepsilon = 0.05$. In contrast, the ID-guided GM allows the reference ID to be perfectly matched. Accordingly, also the KS distances typically improve, from just slightly to even dramatically, as in the case of the closeness, even if not fitted explicitly. The second and third rows represent the same scenario for a 2-dimensional network and the US Power Grid. In such cases, both ABNG and ID-guided ABC fail to even approach the target ID, so that the ID-guided GM sensibly outperforms the other models. As observed in previous scenarios, the KS distances tend to improve, especially in the 2-dimensional case. The green values of the ID and KS were obtained using the ABNG model by explicitly fitting 4 of the 6 properties shown in the KS distance plot (see main text). The error bars on the KS distances are shown in the SI.

of the ABC procedure, we extract 6 different well-known network properties: degree, betweenness, page rank, clustering coefficient, closeness, and eigenvector centrality. These observables are computed on each vertex, defining a set of 6 distributions that is then compared to the corresponding values of the reference network by means of the Kolmogorov–Smirnoff (KS) statistics. The mean of the KS distances for the ensemble of networks of the last generation is reported in the panel on the right. The same plot with error bars is shown in the SI. The picture suggests that while some of the other observable distributions are close to the reference one (KS distance $\lesssim 0.2$), others are not (KS distance $\gtrsim 0.2$). However, we observed in previous cases that, to properly estimate the generative parameters, the average ID of the ensemble of the generated networks has to be very close to the reference ID signature, together with a small variance. It is possible that, by getting closer to the reference ID, the agreement concerning the other properties might improve as well. We will return to this example in the next section, when exploiting our new generative model.

Next, we consider the US Power Grid. To begin with, we start by looking at the ID signature associated with the graphs generated using the optimized parameters reported in[51]. The average over 50 different network realizations (green curve and green circles in the bottom panel of Fig. 4, label: ABNG) shows that the related ID is substantially different from the observed one (black dashed line, label: Observed). In particular, the diameter happens to be much smaller than expected. This is not so surprising, as the suggested parameters were found through an optimization performed by fitting local observables, the ones that also enter as actions in the ABNG model. Accordingly, the meso- and global-scale structures are not properly reproduced. In fact, by examining the associated average KS distances, the quantities used as a target for the fit (betweenness, page rank, clustering coefficient and degree) display very low values ($\lesssim 0.1$), despite the ID curves being very different. Seemingly, eigenvector and closeness centrality measures as well present very high KS statistics.

The blue curve in the bottom left panel of Fig. 4 (label: ID-guided ABC) is the average ensemble ID to which the ID-guided ABC converged after 10 generations, and it represents the closest ID signature that we could reach with this procedure. Even if we manage to find an ID that is consistently closer to the target, the result is again far from satisfactory. Indeed, the obtained diameter is still too low, meaning that the simulated networks are too compact. At the same time, local observables have not consistently worsened (with the exception of the local clustering). This behavior has its origin in the actions used to grow the network. Those are, in fact, based on intuitive mechanisms that leverage on local and neighborhood properties. None of them is explicitly built to enlarge the graph's diameter or enforce global properties. This is a paradigmatic example of the limitations of even the most advanced generative models: their high flexibility allows to sample a wide ensemble of different networks that can reproduce certain network properties. However, other properties, especially of large-diameter networks, are practically impossible to obtain. This observation prompted us to attempt to use a different generative mechanistic model.

Matching the ID curve during the network generation

The reason behind the failure of ID-guided ABC is that the distance matrix—on which ID estimations are based—is a very complicated function of the edges, so that the simple addition of a single edge can dramatically change such a matrix. As a consequence, building intuitive and understandable actions that add edges without compromising the global network structure is far from trivial. For instance, it is very difficult avoid creating bridges/shortcuts among far vertices, whose addition dramatically shortens all distances. To address this problem, we exploit a generative process in which the ID curve is built dynamically, accepting only those moves bringing the ID of the network under construction closer to the observed one (see section "Methods"). To assess the validity of our methodology, we started by applying our ID-guided generative model (GM) to the subset of the Facebook company network that was already presented in the previous section. According to the orange curve and KS distances in the first row of Fig. 4, the ID is now within the desired threshold of $\varepsilon = 0.05$ and the associated KS distances are typically comparable or lower, with a neat improvement especially for closeness, where the mean decreases from 0.3 to 0.05. The worst reproduced property is then the local clustering, with a median KS distance of 0.2.

The next step consists of dealing with large-diameter networks. To this end, we start by analysing an artificial network built on points embedded in a metric space, in the same fashion as those presented in the Introduction (see section "Introduction" and Fig. 1). Once the reference network was created, we applied the ABNG algorithm, ID-guided ABC (again, using ABNG as a generative model) and the ID-guided GM approach. The results are shown in the second row of Fig. 4. Similar to what occurred in the previous example of the US power grid, both ABNG and ID-guided ABC models do not provide a combination of parameters that allows to satisfactorily reproduce the ID signature (green and blue curves). Conversely, the ID-guided GM matches the reference ID within just 3 generations (orange curve). The set of typical network observables are fairly well matched (KS distance $\lesssim 0.2$) apart from the clustering coefficient (KS dist ~ 0.3) and the eigenvector centrality (KS dist ~ 0.5). However, the discrepancy for the latter property is not completely unexpected, as such an observable can display very different distributions (and thus large KS distance), even for networks produced from the same generative model using the same parameters. See SI for a discussion and some examples.

We finally applied the ID-guided GM to the US power grid network. Just three generations of ABC sampling were sufficient to reach an optimal agreement between the observed ID signature and the signature associated with the simulated networks, making the whole experiment less demanding than the one employing the pure ID-guided ABC previously described. In particular, in the former case, the moves involving the addition of edges have an average acceptance rate of order 0.2, while the fractions of accepted graphs were in the range 0.17-0.3. Notably, for the last generation of the ID-guided ABC, the fraction of accepted graphs is of order 10^{-3}, if not lower. The results are reported in orange in the third row of Fig. 4. The posteriors are quite stable along generations, as the target ID signature is reached gradually during the network generation. Still, if we use only random

moves to generate the network, one would need on average 150k Metropolis-Hastings steps to achieve the target number of edges. Conversely, if the links are added according to some specific interpretable rules (similar to those provided by ABNG), as it occurs in our algorithm, only 30k steps are needed. This means that providing structured rules to add the links is still meaningful for increasing the acceptance rate. Very interestingly, apart from the eigenvector centrality, all other measures taken into account display relatively low KS distances from the observed ones. This means that by enforcing the ID signature, one is effectively imposing some restraints that meaningfully affect the local structure, which is fairly well reproduced.

Discussion and conclusions

In this paper, we present a procedure to compute the intrinsic dimension (ID) of unweighted networks by leveraging information from the network at different scales. We then employ the ID as a summary statistic in an Approximate Bayesian Computation (ABC) framework to fit mechanistic generative models. This approach allows for parameter estimation and uncertainty quantification, which significantly contributes to the field. The method can be readily extended to weighted networks with no modification, except that the ID should be estimated with approaches which assume that the distances are real numbers.

We first noted that the ID signature is a powerful observable that allows for the calibration of model parameters. However, we observed that the most advanced and flexible models available in the literature could not correctly reproduce real-world networks' ID signatures, especially when their diameter is large. To address this issue, we developed a new generative process that dynamically controls the ID signature and allows for the satisfactory reproduction of the ID in real-world networks.

As a byproduct, we observed that artificial networks generated through our algorithm show a local structure which is qualitatively comparable with the observed one. Indeed, the probability distribution of local observables, such as the degree, obtained on artificial networks, are very similar to those observed on the target network. Remarkably, consistency is achieved even if these local observables are not included explicitly in the summary statistics used as a target in the ABC approach. In future works, it would be interesting to compare these results with those obtained for networks sampled from statistical models such as ERGM, where the ID signatures at each radius are used as summary statistics to find the vector of model parameters that define the ensemble from which networks are then sampled.

In fact, the choice of which statistics to consider ultimately depends on whether the analysis aims to focus on small-scale details or to understand the overall properties of the network. In this context, we believe the ID signature and the *dk*-statistics[47] can be effectively used together to achieve a more accurate and comprehensive network representation. The *dk* model's ability to accurately replicate local properties complements the ID signature's strength in capturing the meso and global structure.

A possible application of the algorithm could be in network repair/correction: the ID signature would provide a good target to regenerate the broken connections without changing the large scale connectivity of a graph that is only partially known. To prove this statement we run a simple exploratory experiment. We compare the ID signatures averaged over 50 copies of the analysed graphs with a given fraction α of randomly chosen edges removed or added. The results are shown in Fig.5. The FB-Companies' network shows a variation of the ID signature that depends smoothly on α, and is basically "symmetric" under edge addition or removal. This means that, in this network, adding or removing edges does not significantly change the ID signature. Conversely, the ID signature of the 2-dimensional geometric network and of the US power grid are sensitive to link addition and removal in very different ways. Link removal results in very minor variations in the ID signature, as in the case of the FB companies network. Instead, the addition of random links significantly changes the ID signature. We conclude that the ID signature is generally robust under edge removal, even in networks with a large diameter. Additionally, it can detect sudden topological changes when a small number of critical "wrong" edges is added. We plan to further explore the usage of the ID signature as a topological feature in link prediction[57].

Furthermore, our ID estimator could facilitate the task of network representation. The maximum value of the ID could be used as an ansatz for the dimension in which the network should be represented. Moreover, the ID curve can be used as a benchmark to check if the network representation is consistent with the original

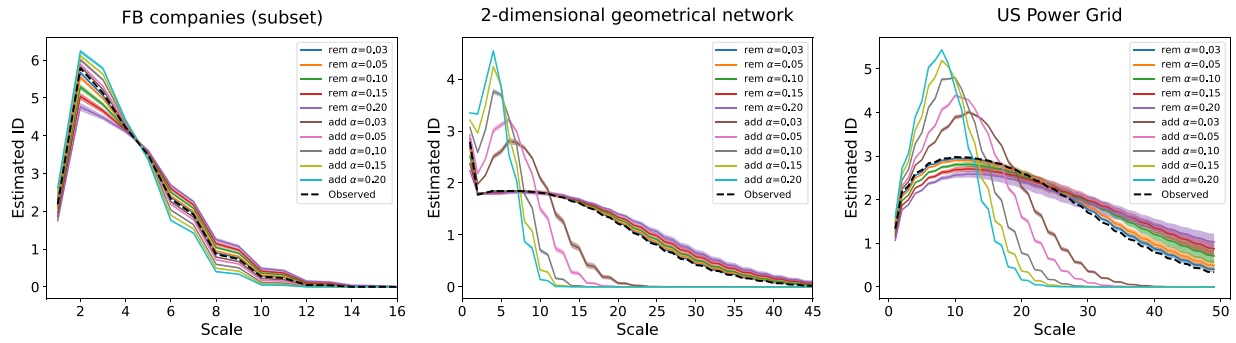

Figure 5. The ID signature is robust under edge removal and sensitive to edge addition. The panels show the ID signatures for ensembles of 50 networks obtained from the original one by the removal or addition of a fraction α of edges.

network. Seemingly, even the complicated task of network lattice embedding[58] can exploit the ID signature as a solid starting point.

Data availability

The software needed to compute the ID for discrete spaces in contained in the DADApy package[59], accessible from here. The extension to networks and the routines used to find the results of the present paper can be found here: https://github.com/imacocco/ABC-NET. The package calls network libraries (Graph-tool[60] and NetworkX[61]) that perform the shortest path calculation and extract other relevant, graph-related observables. All the real-world networks can be found in https://networkrepository.com.

Received: 17 May 2024; Accepted: 19 July 2024
Published online: 01 August 2024

References

1. Bilgin, C. C. & Yener, B. Dynamic network evolution: Models, clustering, anomaly detection. *IEEE Netw.* **1** (2006).
2. Barrat, A., Barthelemy, M. & Vespignani, A. *Dynamical Processes on Complex Networks* (Cambridge University Press, 2008).
3. Costa, L. D. F. *et al.* Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Adv. Phys.* **60**, 329–412 (2011).
4. Latora, V., Nicosia, V. & Russo, G. *Complex Networks: Principles, Methods and Applications* (Cambridge University Press, 2017).
5. Barrat, A., Barthélemy, M. & Vespignani, A. Weighted evolving networks: Coupling topology and weight dynamics. *Phys. Rev. Lett.* **92**, 228701 (2004).
6. Bianconi, G., Darst, R. K., Iacovacci, J. & Fortunato, S. Triadic closure as a basic generating mechanism of communities in complex networks. *Phys. Rev. E* **90**, 042806 (2014).
7. Chakrabarti, D. & Faloutsos, C. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv. (CSUR)* **38**, 2 (2006).
8. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C. & Ghahramani, Z. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.* **11**, 985–1042 (2010).
9. Anderson, C. J., Wasserman, S. & Crouch, B. A p* primer: Logit models for social networks. *Soc. Netw.* **21**, 37–66 (1999).
10. Anand, K. & Bianconi, G. Entropy measures for networks: Toward an information theory of complex topologies. *Phys. Rev. E* **80**, 045102 (2009).
11. Schieber, T. A. *et al.* Quantification of network structural dissimilarities. *Nat. Commun.* **8**, 13928 (2017).
12. Tantardini, M., Ieva, F., Tajoli, L. & Piccardi, C. Comparing methods for comparing networks. *Sci. Rep.* **9**, 17557 (2019).
13. Bagrow, J. P. & Bollt, E. M. An information-theoretic, all-scales approach to comparing networks. *Appl. Netw. Sci.* **4**, 1–15 (2019).
14. Wills, P. & Meyer, F. G. Metrics for graph comparison: A practitioner's guide. *PLoS One* **15**, e0228728 (2020).
15. Solorio-Fernández, S., Carrasco-Ochoa, J. A. & Martínez-Trinidad, J. F. A review of unsupervised feature selection methods. *Artif. Intell. Rev.* **53**, 907–948 (2020).
16. Jović, A., Brkić, K. & Bogunović, N. *A Review of Feature Selection Methods with Applications* 1200–1205 (IEEE, 2015).
17. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
18. Csányi, G. & Szendrői, B. Fractal-small-world dichotomy in real-world networks. *Phys. Rev. E* **70**, 016122. https://doi.org/10.1103/PhysRevE.70.016122 (2004).
19. Gastner, M. T. & Newman, M. E. The spatial structure of networks. *Eur. Phys. J. B Condens. Matter Complex Syst.* **49**, 247–252 (2006).
20. Daqing, L., Kosmidis, K., Bunde, A. & Havlin, S. Dimension of spatially embedded networks. *Nat. Phys.* **7**, 481–484 (2011).
21. Silva, F. N. & Costa, L. D. F. Local dimension of complex networks. *arXiv preprint*arXiv:1209.2476 *(2012).*
22. Song, C., Havlin, S. & Makse, H. A. Self-similarity of complex networks. *Nature* **433**, 392–395 (2005).
23. Boguna, M. *et al.* Network geometry. *Nat. Rev. Phys.* **3**, 114–135 (2021).
24. Gallos, L. K., Song, C. & Makse, H. A. A review of fractality and self-similarity in complex networks. *Physica A Stat. Mech. Appl.* **386**, 686–691 (2007).
25. Wen, T. & Cheong, K. H. The fractal dimension of complex networks: A review. *Inf. Fusion* **73**, 87–102 (2021).
26. Rosenberg, E. *Fractal Dimensions of Networks* Vol. 1 (Springer, 2020).
27. Falconer, K. *Fractal Geometry: Mathematical Foundations and Applications* (Wiley, 2004).
28. Schneider, C. M., Kesselring, T. A., Andrade, J. S. Jr. & Herrmann, H. J. Box-covering algorithm for fractal dimension of complex networks. *Phys. Rev. E* **86**, 016707 (2012).
29. Grassberger, P. & Procaccia, I. Characterization of strange attractors. *Phys. Rev. Lett.* **50**, 346 (1983).
30. Lacasa, L. & Gómez-Gardenes, J. Correlation dimension of complex networks. *Phys. Rev. Lett.* **110**, 168703 (2013).
31. Macocco, I., Glielmo, A., Grilli, J. & Laio, A. Intrinsic dimension estimation for discrete metrics. *Phys. Rev. Lett.* **130**, 067401. https://doi.org/10.1103/PhysRevLett.130.067401 (2023).
32. Marin, J.-M., Pudlo, P., Robert, C. P. & Ryder, R. J. Approximate Bayesian computational methods. *Stat. Comput.* **22**, 1167–1180 (2012).
33. Sunnåker, M. *et al.* Approximate Bayesian computation. *PLoS Comput. Biol.* **9**, e1002803 (2013).
34. Di Noia, A., Macocco, I., Glielmo, A., Laio, A. & Mira, A. Robust intrinsic dimension estimation via optimal neighbourhood identification. Under Review.
35. Eugène ehrhart—publications 1947–1996. http://icps.u-strasbg.fr/~clauss/Ehrhart_pub.html. Accessed: 2022-03-25.
36. Beck, M. & Robins, S. Computing the continuous discretely: Integer-point enumeration in polyhedra. *Choice Rev. Online* **45**, 45-0923. https://doi.org/10.5860/choice.45-0923 (2007).
37. Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. & Robert, C. P. Adaptive approximate Bayesian computation. *Biometrika* **96**, 983–990 (2009).
38. Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202 (2009).
39. Fay, D., Moore, A. W., Brown, K., Filosi, M. & Jurman, G. Graph metrics as summary statistics for approximate Bayesian computation with application to network model parameter estimation. *J. Complex Netw.* **3**, 52–83 (2015).
40. Chen, S., Mira, A. & Onnela, J.-P. Flexible model selection for mechanistic network models. *J. Complex Netw.* **8**, cnz024 (2020).
41. Raynal, L., Chen, S., Mira, A. & Onnela, J.-P. Scalable approximate Bayesian computation for growing network models via extrapolated and sampled summaries. *Bayesian Anal.* **17**, 165–192 (2022).
42. Sisson, S. A., Fan, Y. & Beaumont, M. A. Overview of abc. *Handbook of Approximate Bayesian Computation*, 3–54 (2018).
43. Sisson, S. A., Fan, Y. & Tanaka, M. M. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.* **104**, 1760–1765 (2007).
44. Sinitsyn, N., Hengartner, N. & Nemenman, I. Adiabatic coarse-graining and simulations of stochastic biochemical networks. *Proc. Natl. Acad. Sci.* **106**, 10546–10551 (2009).

45. Czyżak, P. & Jaszkiewicz, A. Pareto simulated annealing. In *Multiple Criteria Decision Making: Proceedings of the Twelfth International Conference Hagen (Germany)*, 297–307 (Springer, 1997).
46. Schälte, Y., Klinger, E., Alamoudi, E. & Hasenauer, J. pyabc: Efficient and robust easy-to-use approximate Bayesian computation. *J. Open Source Softw.* **7**, 4304. https://doi.org/10.21105/joss.04304 (2022).
47. Orsini, C. *et al.* Quantifying randomness in real networks. *Nat. Commun.* **6**, 8627 (2015).
48. Jamakovic, A., Mahadevan, P., Vahdat, A., Boguná, M. & Krioukov, D. How small are building blocks of complex networks. *arXiv preprint*arXiv:0908.1143 *(2009)*.
49. Rossi, R. A. & Ahmed, N. K. The network data repository with interactive graph analytics and visualization. In *AAAI* (2015).
50. Erdős, P. *et al.* On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960).
51. Arora, V. & Ventresca, M. Action-based modeling of complex networks. *Sci. Rep.* **7**, 1–10 (2017).
52. Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629–4632. https://doi.org/10.1103/PhysRevLett.85.4629 (2000).
53. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
54. Condon, A. & Karp, R. M. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms* **18**, 116–140 (2001).
55. Chung, F. & Lu, L. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci.* **99**, 15879–15882 (2002).
56. Chung, F. & Lu, L. Connected components in random graphs with given expected degree sequences. *Ann. Combin.* **6**, 125–145 (2002).
57. Ran, Y., Xu, X.-K. & Jia, T. The maximum capability of a topological feature in link prediction. *PNAS Nexus* **3**, pgae113 (2024).
58. Eppstein, D. The lattice dimension of a graph. *Eur. J. Combin.* **26**, 585–592 (2005).
59. Glielmo, A. *et al.* Dadapy: Distance-based analysis of data-manifolds in python. *Patterns* **3**, 100589. https://doi.org/10.1016/j.patter.2022.100589 (2022).
60. Peixoto, T. P. The graph-tool python library. *figshare*https://doi.org/10.6084/m9.figshare.1164194 *(2014)*.
61. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference* (eds Varoquaux, G. *et al.*) 11–15 (Pasadena, 2008).
62. Silverman, B. W. *Density Estimation for Statistics and Data Analysis* (Routledge, 2018).
63. Filippi, S., Barnes, C. P., Cornebise, J. & Stumpf, M. P. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat. Appl. Genet. Mol. Biol.* **12**, 87–107 (2013).

Acknowledgements

We acknowledge Antonio di Noia for the precious help and long discussions. We also thanks Conor Hassan and Viplove Arora for tips and suggestions in using the code and writing the manuscript. Antonietta Mira acknowledges financial support from SNSF grant 200021_208249 "Feature Learning for Bayesian Inference".

Author contributions

I.M., A.M. and A.L. designed the research and wrote the paper; I.M. performed the research; A.M. and A.L. contributed to perform the research.

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/s41598-024-68113-3.

Correspondence and requests for materials should be addressed to A.L.

Reprints and permissions information is available at www.nature.com/reprints.

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