Inferring pattern generators on networks

Piotr Nyczka, Marc-Thorsten Hütt, Annick Lesne

Inferring pattern generators on networks

Piotr Nyczka, Marc-Thorsten Hütt

Department of Life Sciences and Chemistry, Jacobs University, D-28759 Bremen, Germany

Annick Lesne

Sorbonne Université, CNRS, Laboratoire de Physique Théorique de la Matière Condensée, LPTMC, F-75252, Paris, France

[a]*Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, F-34293, Montpellier, France*

Abstract

Given a pattern on a network, i.e. a subset of nodes, can we assess, whether they are randomly distributed on the network or have been generated in a systematic fashion following the network architecture? This question is at the core of network-based data analyses across a range of disciplines — from incidents of infection in social networks to sets of differentially expressed genes in biological networks. Here we introduce generic 'pattern generators' based on an Eden growth model. We assess the capacity of different pattern measures like connectivity, edge density or various average distances, to infer the parameters of the generator from the observed patterns. Some measures perform consistently better than others in inferring the underlying pattern generator, while the best performing measures depend on the global topology of the underlying network. Moreover, we show that pattern generator inference remains possible in case of limited visibility of the patterns.

Keywords: Patterns; Network clusters; Teleportation random walks; Eden model; Parametric inference; Mutual information

Published in Physica A, 566: 125631 (2021). Special Issue: 'Interdisciplinary applications of statistical physics in memory of Professor Dietrich Stauffer'.

Email address: annick.lesne@sorbonne-universite.fr (Annick Lesne)

Highlights

- We investigate various Eden model-based pattern generators on networks.
- Generator parameters can be inferred from characteristics of the observed patterns.
- The quality of the inference depends on which pattern characteristic is measured.
- Which measure is the most suitable depends on the underlying network global topology.
- The inference remains possible in case of degraded observation of the patterns.

1. Introduction

Patterns on networks are a natural object of investigation. Networks can here be real objects, for instance neural networks, or the representation of a substrate, ranging from the discretization of real space into lattices to complex networks whose edges indicate some interaction or relationship between entities each corresponding to a node of the network. As in percolation studies [1], some nodes can be singled out (henceforth termed 'colored') based on their binary (or binarized) state, e.g. occupied vs. empty. Often, the binary state of the nodes results from some dynamical process on the network, e.g. infected nodes in virus or rumor propagation, active nodes in a neural network or expressed genes in a biological network, or more generally nodes visited by some transport process [2] or contact process [3]. In all cases, network patterns are formed by clusters (i.e. connected subsets) of colored nodes.

The question addressed in the present study is the possibility to infer the level of randomness of a network pattern from some observable quantity characterizing the pattern. For a diffusion process, it is well known that the features of the clusters of visited nodes are sensitive both to the parameters of the transport

process and the underlying network topology [4, 5]. We will thus consider various networks, ranging from lattices to complex networks such as Erdős-Rényi random graphs (ER)[6], Watts-Strogatz small worlds (SW) [7] and Barabasi-Albert scale-free graphs (BA) [8]. We will investigate patterns generated on the networks by a class of processes combining an Eden growth model [9] and teleportation events [10], in which a new seed node is chosen at random. Such a dual model could mimic the features of biological transport [11, 12]. Its main interest here is to include a tunable level of randomness, namely the probability $(1 - p)$ of a teleportation event. For $p = 0$, the generated pattern is fully random, whereas the standard Eden model is recovered for $p = 1$.

Quantifying the amount of clustering (or other indicators of a non-random distribution) of colored nodes in a larger graph is a key task for example in any network-based analysis of biological data in Systems Biology and Systems Medicine [13]. In these cases, typical graphs are protein interaction networks [14, 15], metabolic networks [16, 17] or signaling networks [18]. Examples of these studies include analyzing the distribution of disease-associated genes in protein interaction networks [19] and other biological networks [20], the network-based enhancement of biological signals derived from high-throughput data [21] and the network-based interpretation of gene expression patterns [22, 23, 24].

Relatedly, interpreting incidence patterns of an infectious disease can also be approached by comparing the patterns to a given underlying network architecture [25]. Similar questions also arise in the context of Computational Neuroscience, when activity patterns of cortical areas (often denoted as 'functional connectivity') are compared with the underlying anatomy of cortical area interactions (or 'structural connectivity'); see, e.g., [26, 27]. The analysis and interpretation of the correlations between structural connectivity and functional connectivity also extends to a range of other disciplines [28].

Our study will be based on extensive simulation, in the spirit of the pioneering work of Dietrich Stauffer [4, 11, 29]. Given the above-described general class of pattern generators (combinations of Eden growth model and random hops), we will investigate, if it is possible to infer their parameters from an

appropriate measure of the patterns generated on a given underlying network, even when they are observed with incomplete information. More precisely, we want (1) to infer the parameters of the pattern generator, (2) to determine the dependence of the inference result on the topology of the underlying network, and (3) to assess the performance of the various observables (pattern measures) as an input of the inference process. In contrast to many studies of transport processes on networks, we will not consider here the temporal evolution of the pattern features [2]. We will focus on the relationship between the parameters of the pattern generators, the observable features of the generated patterns, and the underlying network topology.

2. Pattern generation

We first describe how the sets of colored nodes, that we denote as 'patterns', are generated on the considered graph $\Gamma(E, V)$, where V is the set of nodes and E the set of edges connecting them. For this purpose a modified Eden growth model with teleportation is employed, according to the following scheme:

1. Set values of all nodes to 0.
2. Randomly select an uncolored node (having value 0) of the graph and color it, i.e., set its value to 1.
3. (a) With probability p chose an uncolored neighbor of the current colored cluster and color it.
   (b) With probability $1 - p$ go to point 2 and set a new cluster.

The algorithm stops when the desired number n of vertices is colored. The parameter p controls the 'cohesion' or compactness of the colored nodes, and will henceforth be called the 'cohesion parameter'. The lower p the more fragmented the patterns will be. For $p = 0$ colored nodes are distributed evenly in a random fashion, while for $p = 1$ there is only one big cluster of colored nodes. Analogous procedure, but without teleportation (i.e. $p = 1$) is known as Eden model [9].

To each pattern corresponds an induced subgraph $\Gamma_c(E_c, V_c) \subseteq \Gamma(E, V)$, where V_c is the subset of colored nodes and E_c is the subset of edges connecting

two colored nodes. Note that we are interested in the generated patterns Γ_c, not in the evolution of colored clusters as a function of time [2].

In addition to the size of the set of nodes forming the pattern, which is is kept constant (equal to 50) throughout our investigation, the pattern generators have the following parameters:

- Cohesion parameter p, with $1-p$ being the teleportation probability;
- Environmental range e, which is the radius of the neighborhood of a colored node. In practice, e it is the thickness of the layer surrounding the colored clusters, in which Eden growth could occur. For $e = 1$, we include nearest neighbors in the layer, for $e = 2$ next-to-nearest neighbors and so on.
- Visibility q, with $1-q$ being the probability of deleting a node from the pattern (i.e. the generated set of colored nodes). Denoting n the number of colored nodes, for $q < 1$ we generally observe less than n nodes. Denoting n' be the number of observed nodes, we have $\langle n' \rangle = qn$.

Figure 1 illustrates on a small 2D lattice the diversity of colored node sets (i.e. 'patterns') created by our pattern generator, as well as the influence of its key parameters.

3. Inference of the pattern generator

The question we ask is very simple: How far from randomness is the given pattern? What is needed is a computational resource converting a given network and the pattern of colored nodes into a single number. This single number should be seen as a quantifier for the (lack of) randomness of the pattern. In other words, it should tell, whether the given pattern is random, or not, and quantify this deviation from randomness with the highest precision possible. In the present context, pattern randomness is directly related to the value of the cohesion parameter p of the generator, with $p = 0$ corresponding to full randomness (see top right pattern on Fig. 1) whereas $p = 1$ corresponds to the

Eden model (see top left pattern on Fig. 1). Our goal will be to infer the value of p from the observation of the colored patterns.

3.1. Definition of pattern measures

In the following, we introduce three types of pattern quantifiers: (i) measures based on edge density; (ii) measures based on cluster count; (iii) measures based on the distances between colored nodes. Each measure summarizes a pattern on the network in just a single number. The measure will then be used to infer the parameter p of the pattern generator. The observed value of the measure for a given pattern will be termed an 'observable'.

3.1.1. Edge density-based measures X_e

The simplest type of measures considered here are those relying on basic statistics of the colored subgraph $\Gamma_c(E_c, V_c)$ (comprising $n = |V_c|$ nodes) extracted from the original graph $\Gamma(E, V)$. After extraction the measures involve solely the subgraph Γ_c with no relation to the original graph Γ.

Connectivity. It is the fraction of non-isolated nodes, i.e. of nonzero degree in the induced subgraph Γ_c:

$$X_e^c = \frac{1}{|V_c|} \sum_{j \in V_c} \{1 | k_j > 0\} = \frac{1}{n} \sum_{j \in V_c} \{1 | k_j > 0\} \ . \tag{1}$$

Edge density. It is the fraction of edges actually present in the induced subgraph Γ_c, compared to the maximal number of edges that could be established between the n nodes of Γ_c:

$$X_e^d = \frac{2|E_c|}{|V_c|(|V_c| - 1)} = \frac{2|E_c|}{n(n-1)} \ . \tag{2}$$

Average clustering coefficient. Denoting C_j the clustering coefficient in the induced subgraph Γ_c of a node j of the subgraph, its average is:

$$X_e^{cl} = \frac{1}{|V_c|} \sum_{j \in V_c} C_j = \frac{1}{n} \sum_{j \in V_c} C_j \ . \tag{3}$$

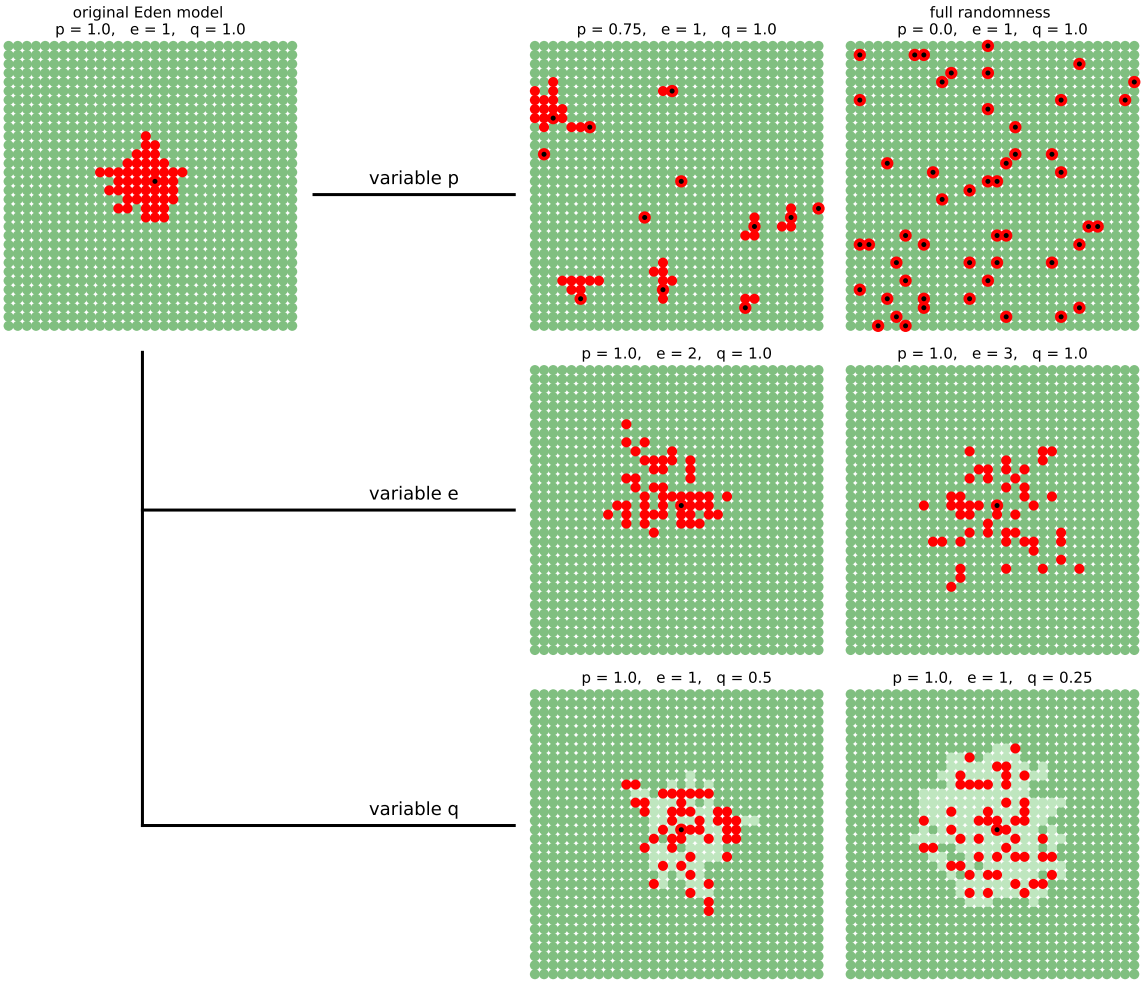


Figure 1: (color online) 2D lattice illustration of the variety of patterns (all comprising 50 nodes) obtained for different values of the generator parameters. The pattern on top left has been obtained with our default setting (p=1, e=1, q=1), corresponding to the original Eden model. Patterns on the first row have been obtained with higher teleportation probabilities $(1 - p)$, increasing along the row (i.e., decreasing values of the parameter p controlling the cohesion of the pattern). The pattern on top right illustrates the fully random patterns obtained for parameters (p=1, e=1, q=1). The second row presents patterns obtained at increasing values of the environmental range e. Along the third row, the visibility q decreases. Red + black: colored nodes (red: nodes selected within the neighborhood of the already grown clusters, according to the Eden growth model; black: nodes selected randomly due to a teleportation event). Green and light green: nodes *not* in the set of colored nodes (light green: nodes previously colored, which have been discarded from the pattern due to a limited visibility q). Note that the distinction between red and black, as well as the distinction between green and light green, are only used for visualization, in order to better illustrate the features of the pattern generator; they do not affect the analysis.

3.1.2. Cluster count X_n

Another type of measures are those relying on counting the number of connected clusters in the colored subgraph Γ_c. Here we used three variants: X_n^1 is the number of connected clusters within the induced subgraph Γ_c; X_n^2 is the analogous quantity calculated for the pattern where a direct link is established between two colored nodes if they are related by paths of length 1 or 2 on the original network, respectively paths of length up to 3 for X_n^3.

3.1.3. Distance based X_r

We have also benchmarked several distance-based measures. Their definition follows a common framework with tunable elements, which yields a large diversity of measures (overall 60 possible combinations):

$$X_r^{\alpha,\beta,y} = \frac{1}{m}\sum_j^m \omega_j r_j^y \ , \tag{4}$$

where ω_j is a weight, r_j a distance, and $y \in \{-4,-3,-2,-1,-\frac{1}{2},\frac{1}{2},1,2,3,4\}$ an exponent. The superscripts α and β are mere labels detailed below, defining in particular the set of m indices over which the sum runs. Distances are calculated in two ways:

1. For each pair j of colored nodes g and h, we compute a distance r_j equal to the shortest path between g and h. The sum here runs over the m pairs of colored nodes. The corresponding measure will be labeled by a superscript $\alpha = A$.
2. For each colored node j, we calculate the distance r_j to the nearest colored node. The sum here runs over the $m = n$ colored nodes. The corresponding measure will be labeled by a superscript $\alpha = a$.

Additionally there are three choices for the weight ω_j:

1. Each distance r_j is endowed with an equal weight $\omega_j = 1$; the corresponding measure will be labeled by a superscript $\beta = 0$.
2. for distances calculated following the way (1) above:

- $\omega_j \equiv k_g k_h$, where j labels the pair of colored nodes (g, h) and k is the degree of a node; this choice will be labeled $\beta = \uparrow$.
- $\omega_j \equiv \frac{1}{k_g k_h}$, where j labels the pair of colored nodes (g, h) and k is the degree of a node; this choice will be labeled $\beta = \downarrow$.

3. For distances calculated following the way (2) above:
   - $\omega_j = k_j$, where k_j is degree of the colored node j; this choice will be labeled $\beta = \uparrow$.
   - $\beta = \downarrow$, $\omega_j = \frac{1}{k_j}$, where k_j is degree of the colored node j; this choice will be labeled $\beta = \downarrow$.

As smaller y enhances the contribution of closer distances, there are two limit cases where two distance-based measures recover connectivity-based measures:

$$\lim_{y \to -\infty} X_r^{A,0,y} = X_e^d \,, \tag{5}$$

$$\lim_{y \to -\infty} X_r^{a,0,y} = X_e^c \,. \tag{6}$$

In our study, we highlighted four specific measures:

1. the connectivity: X_e^c ;
2. the average clustering coefficient: X_e^{cl} (denoted "clustering" on plots);
3. the cluster count of order 1: X_n^1 (denoted"clusters1_n" on plots);
4. one of the distance-based measures: $X_r^{A,0,-1}$ (denoted "A:rˆ-1" on plots).

3.2. Simulations

In order to evaluate the inference power of the above-defined measures, we numerically implemented the pattern generator and quantified the generated patterns with all the measures. The simulation setting is characterized by the network architecture, the value of the cohesion parameter p, the choice of the neighborhood e and the value of the visibility q. In each considered instance, we performed 100 simulation runs to get reasonable datasets. In all the simulations we used networks with $N = 1000$ nodes except for the square lattice where $N = 32 \cdot 32 = 1024$. Networks had different topologies and different average

degree $\langle k \rangle$. For complex networks the environmental range was $e \in \{1, 2, 3\}$ while for lattices $e \in \{1, 2, ..., 10\}$. We also set $\langle n' \rangle = nq = const. = 50$ and used three different combinations of the parameters n and q, namely: ($n = 50$, $q = 1.0$), ($n = 100$, $q = 0.5$) and ($n = 200$, $q = 0.25$). Finally the cohesion parameter p spanned its whole range from 0 to 1 with 0.1 increments.

As the performance of the measures may depend on the range of the generator parameter p to be inferred, we distinguished the whole range ($0 \leq p \leq 1$) from the 'weak signal regime' where the range of the cohesion parameter is restricted to $0 \leq p \leq 0.2$, which produces only diffuse patterns and low observable values. This latter case is particularly important for applications to medical data [30].

3.3. Evaluation of a measure

We focused on the predictive power of a given measure regarding the cohesion parameter p, knowing the other parameters e and q. This parameter p controls the level of randomness of the generated pattern. We used the well-established and widely used notion of mutual information for evaluating a given measure. Indeed the mutual information $I(X; p)$ of two random variables X and p is the quantifier telling us how much we can know about one random variable, knowing the value of the other. In other words it is an alternative for a correlation coefficient, however more precise, especially in the nonlinear cases. The observable X is here a random variable as it comes from a randomly generated pattern. In an inference problem, the cohesion parameter p, which is fixed in a simulation, also becomes a random variable. This idea drives our analysis strategy.

The set of simulation results provides the joint distribution $P_{(X,p)}(X, p|e, q)$ of the considered observable X and the cohesion parameter p, given the parameters e and q. For simplicity, we note in the same way the observable values and the observable considered as a random variable, and we will skip in the notation the explicit mention of e and q. This distribution was reconstructed at fixed values of e and q using eleven bins for X (ranging between the extreme values

observed for X) and eleven bins for the eleven considered values of p (from 0 to 1 by steps of 0.1), except in the weak-signal regime $p \leq 0.2$ where only three bins were used. From the joint probability distribution, we computed the marginal distributions $P_X(X)$ and $P_p(p)$, and the mutual information:

$$I(X;p) = H(X) - H(X|p) = \sum_X \sum_p P_{(X,p)}(X,p) \log\left(\frac{P_{(X,p)}(X,p)}{P_X(X)P_p(p)}\right) , \quad (7)$$

where $H(X)$ is the Shannon entropy (the sum runs over the discrete values of X)

$$H(X) = -\sum_X P(X) \log(P(X)) . \quad (8)$$

This mutual information provides an overall quantification of the reconstructability and predictability of the parameter p from the observation of X, that will be used to rank the measures. The higher $I(X;p)$ the more information about p is given by X. The best performing measures regarding the inference of the cohesion parameter p will be those with the highest $I(X;p)$.

In practical applications (as, e.g., in [23, 24]), it can be helpful to transform the measures X into a z-score with respect to a suitable null model (e.g., randomly drawn nodes), in order to make situations with very different sizes of data sets and networks comparable. Here, replacing the measures by their z-score is useless, as the mutual information is invariant with respect to such a linear transform. The reason for this is straightforward, as I depends only on the probabilities of certain values, not on the values themselves [31].

In the following we evaluate the inference of the cohesion parameter p from the set of measures for different values of the other pattern generator parameters (environmental range e and visibility q, which are fixed in the evaluation of $P_{(X,p)}(X,p)$ and the inference of p).

An illustration of the inference of pattern generator parameters from pattern observables is presented in Figure 2. This figure displays four observables X as a function of the cohesion parameter p for different values of the environmental range e. A range of observed values X, obtained from a large set of simulated patterns, translates into a range of values of p. These plots give an intuition of

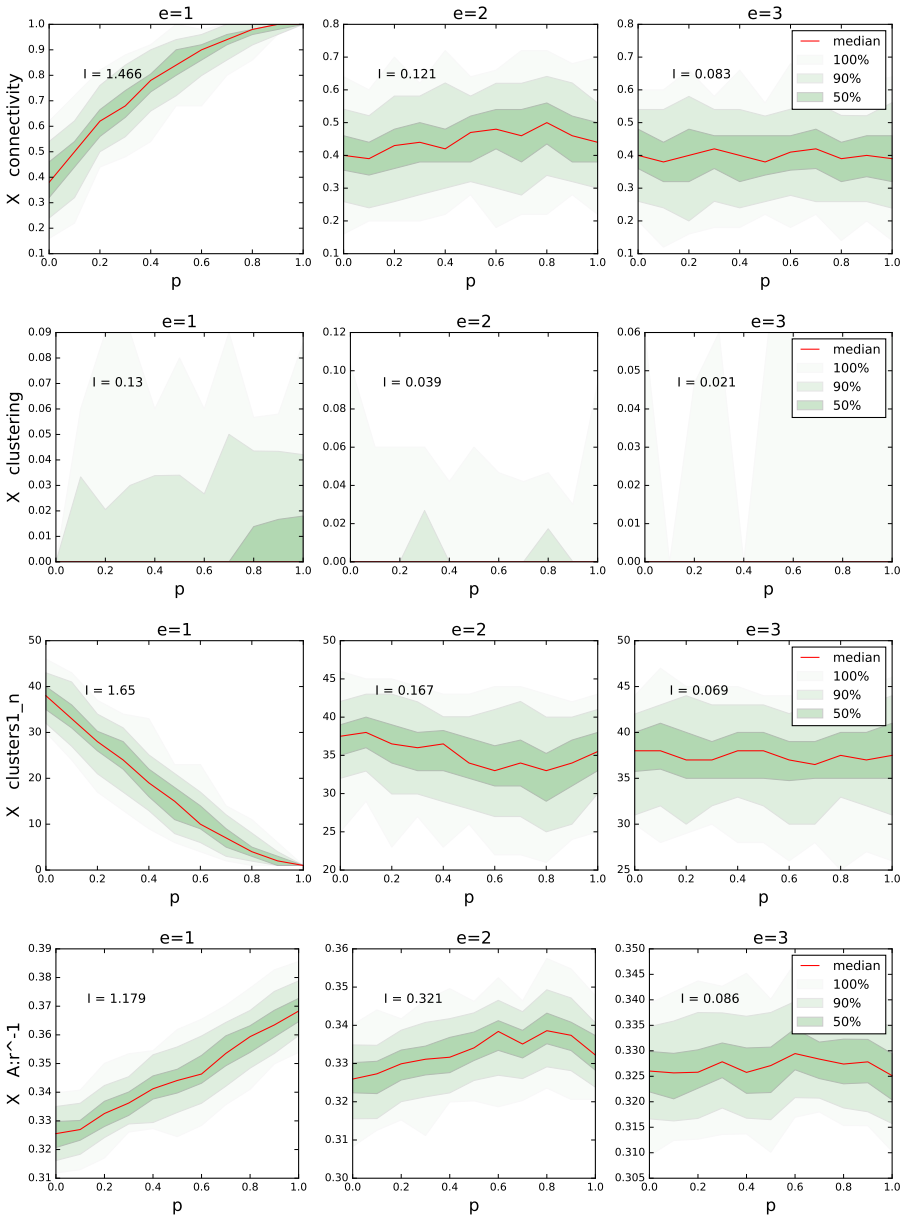


Figure 2: (color online) Inference of the pattern generator parameter p from sets of patterns at fixed values of e and q, illustrated for four of the measures used in our investigation: connectivity (first row), average clustering coefficient (second), cluster count (third row) and a distance-based measure (fourth row). The observables X are shown as a function of the cohesion parameter p for different values of the environmental range e. In each instance, the value of the mutual information $I(X, p)$ (in short I) is indicated. First column: $e = 1$. Second column: $e = 2$. Third column: $e = 3$. All cases correspond to an Erdős-Rényi graph (ER) [6] with $N = 1000$ nodes, $\langle k \rangle = 5$, $n = 50$, $q = 1$, and for each set of parameters 100 samples were simulated.

the sensitivity of an observable X with respect to the variation of the parameter p: When the slope of the plot is steep, even a small variation in p will correspond to a detectable variation in the range of values of X, and could thus be inferred from the observation of X. An overall assessment of the inference quality for each observable and each value of e is provided by the value of the mutual information $I(X, p)$, the higher the better, indicated in each panel of the figure. We see, for example, that for $e = 1$ three measures, the connectivity (first row), the cluster count (third row) and the distance-based measure (fourth row) can be used to infer the parameter p with high accuracy. The quality of this inference decreases rapidly with the environmental range e. For $e = 2$ and $e = 3$, we observe that different values of p could give the same value of the measure, e.g. second column fourth row; in this case, only bounds on the value of p could be derived from the pattern measure. The clustering coefficient (second row), on the other hand, does not allow inference of p even for $e = 1$, due to the low values of the clustering coefficient in the used network.

A first set of results for Erdős-Rényi (ER) graphs is shown in Figure 3, in the form of curves displaying the mutual information $I(X, p)$ as a function of the average degree $\langle k \rangle$ of the network. The full set of observables is shown as gray curves, while the colored curves highlight the four selected measures: connectivity, clustering, cluster count, and one distance-based measure. The curves confirm the visual impression from Figure 2, extend it to a wide range of networks (due to the variation of the average degree) and put the four highlighted observables in the context of the large set of other observables (shown in gray). Generally, a low value of $I(X, p)$, reflecting a poor inference of the parameter p, is observed either in case of a low sensitivity of the measure to variations of p (e.g. third column first row in Fig. 2) or the case when two values of p gives the same value of the measure (e.g. second column fourth row in Fig. 2). Furthermore, Figure 3 distinguishes between the inference across the whole range of p (left column) and the 'weak signal' regime $p \leq 0.2$ (right column), corresponding to diffuse patterns and low observable values. In particular, we see that the connectivity and the cluster count perform well with an

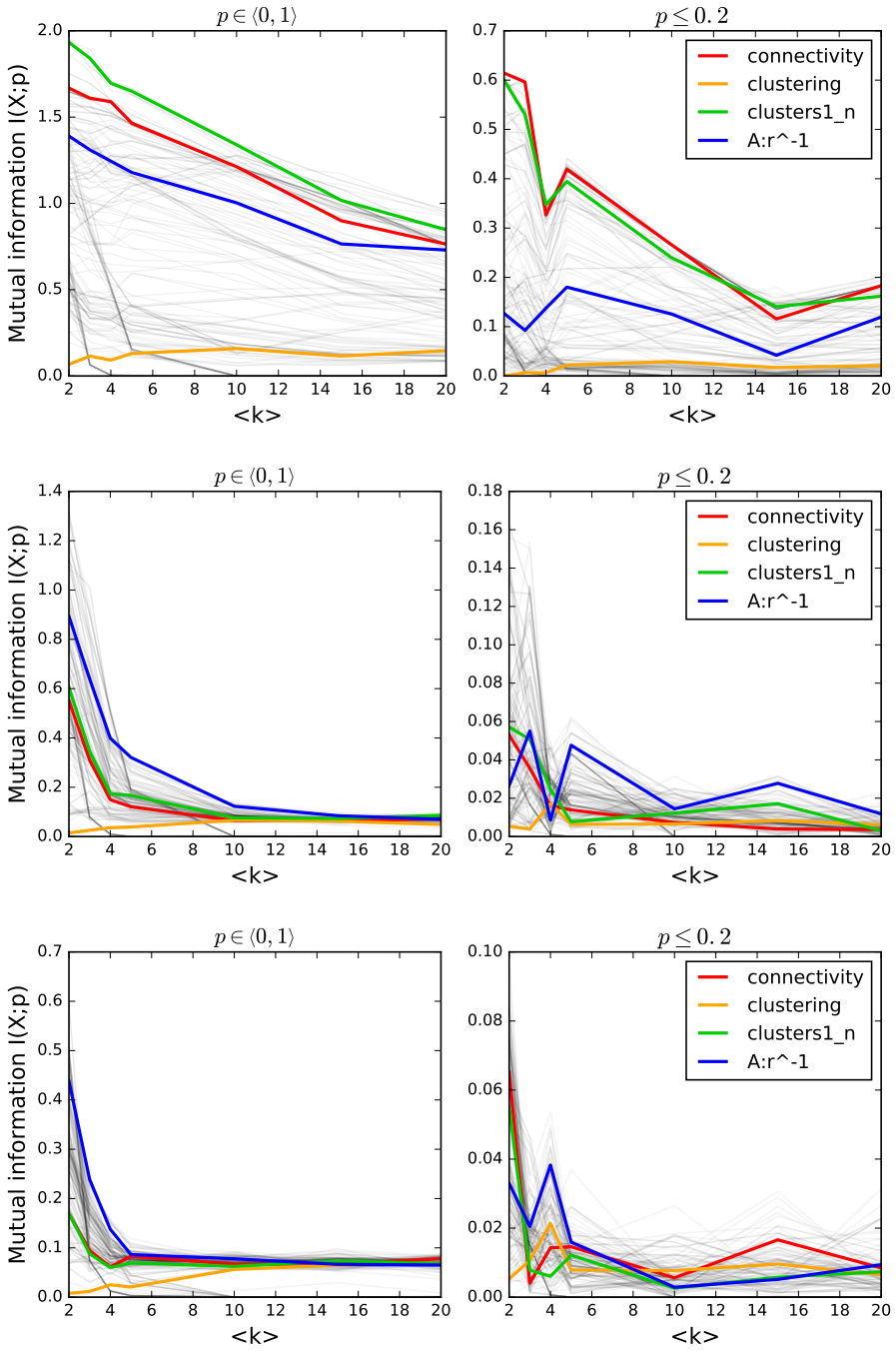


Figure 3: color online) Mutual information (quantifying the performance of a measure for parameter inference) as a function of the average degree of the network for the case of ER graphs. Left column: mutual information evaluated across the whole range of the cohesion parameter p. Right column: mutual information evaluated for small p (yielding diffuse patterns and weak observable values). First row: $e = 1$. Second row: $e = 2$. Third row: $e = 3$. A rapid deterioration of the predictive power is observed for $e > 1$ along with a shuffling of the best performing measures. For $e > 1$ the distance-based measure $X_r^{A,0,-1}$ becomes significantly better than the others. While for $e = 1$ edge density and cluster count perform better overall, for weak signals (second column) connectivity has a slight advantage over them for $\langle k \rangle < 9$.

advantage for the connectivity in the case of weak signals and sparse graphs. With increasing e (second and third row), the representative of a distance-based measure (blue curve) performs best, even though the performance of measures in general decreases substantially with the environmental range e, as expected.

3.4. Different network architectures

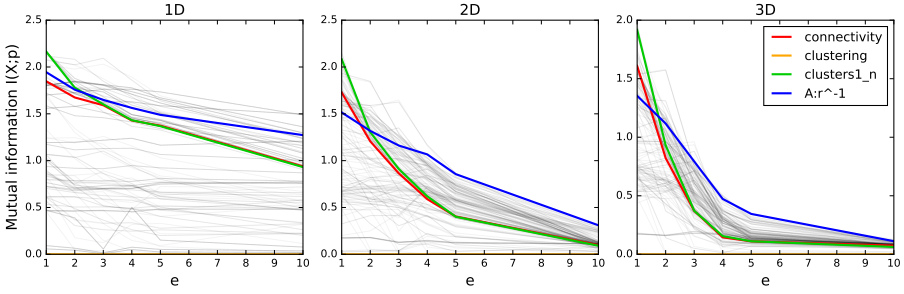
In Figure 4 the performances of the different measures is explored with respect to the global network architecture. The highlighted measures display a similarly satisfactory performance for lattices (first row; same as the first row in Figure 3) and random regular graphs (second row), except for the clustering-based measure, which performs poorly, as in the case of the ER graph discussed before. In contrast, differences appear when considering more complex network architectures, namely small-world graphs (third row) and scale-free graphs (fourth row): There, connectivity and cluster count perform almost equally well for scale-free graphs, while cluster count performs better for small-world graphs. Additional details about small-world graphs are given in Figure S1.

Note that for lattices, results are presented as a function of the environmental range e (rather than $\langle k \rangle$), for two reasons: On lattices, average distances are greater than on complex networks hence we could consider larger values of e, while on the other hand $\langle k \rangle$ values are fixed. For lattices distance-based measures are the most efficient and they perform well even for large e.

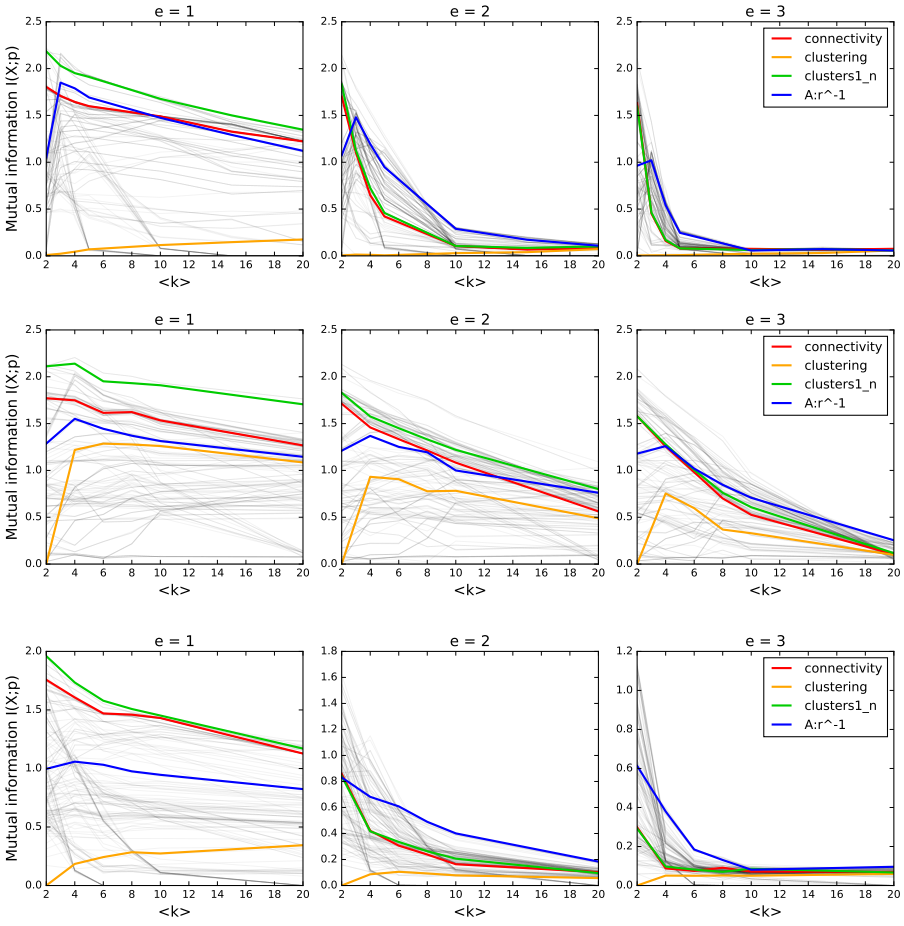
3.5. Accommodating incomplete observation

With decreasing visibility q (recall that q is fixed in the inference procedure), our highlighted measures differ in particular at weak signal strength (low cohesion parameter p, diffuse patterns, weak observable values) with a tendency of the connectivity measure to perform best (see Fig. 5).

Figure S2 offers a more detailed look at lattices, showing the performance in inferring the parameter p of the various measures as a function of the environmental range e for different space dimensions (columns) and different values of the visibility q (rows). It is striking to see that on these network architectures, the distance-based measure is almost not affected by decreasing visibility.



(a) Lattices



(b) Networks

Figure 4: (color online) Same as Figure 3, but for different network architectures. First row: lattices of dimension 1, 2 and 3, as a function of the environmental range e (horizontal axis) Second row: random regular graphs (RR). Third row: small-world graphs (SW) [7]. Fourth row: scale-free graphs (BA) [8]. For the latter three rows, the results are presented as a function of the average degree $\langle k \rangle$. Left column: $e = 1$. Middle column: $e = 2$. Right column: $e = 3$.

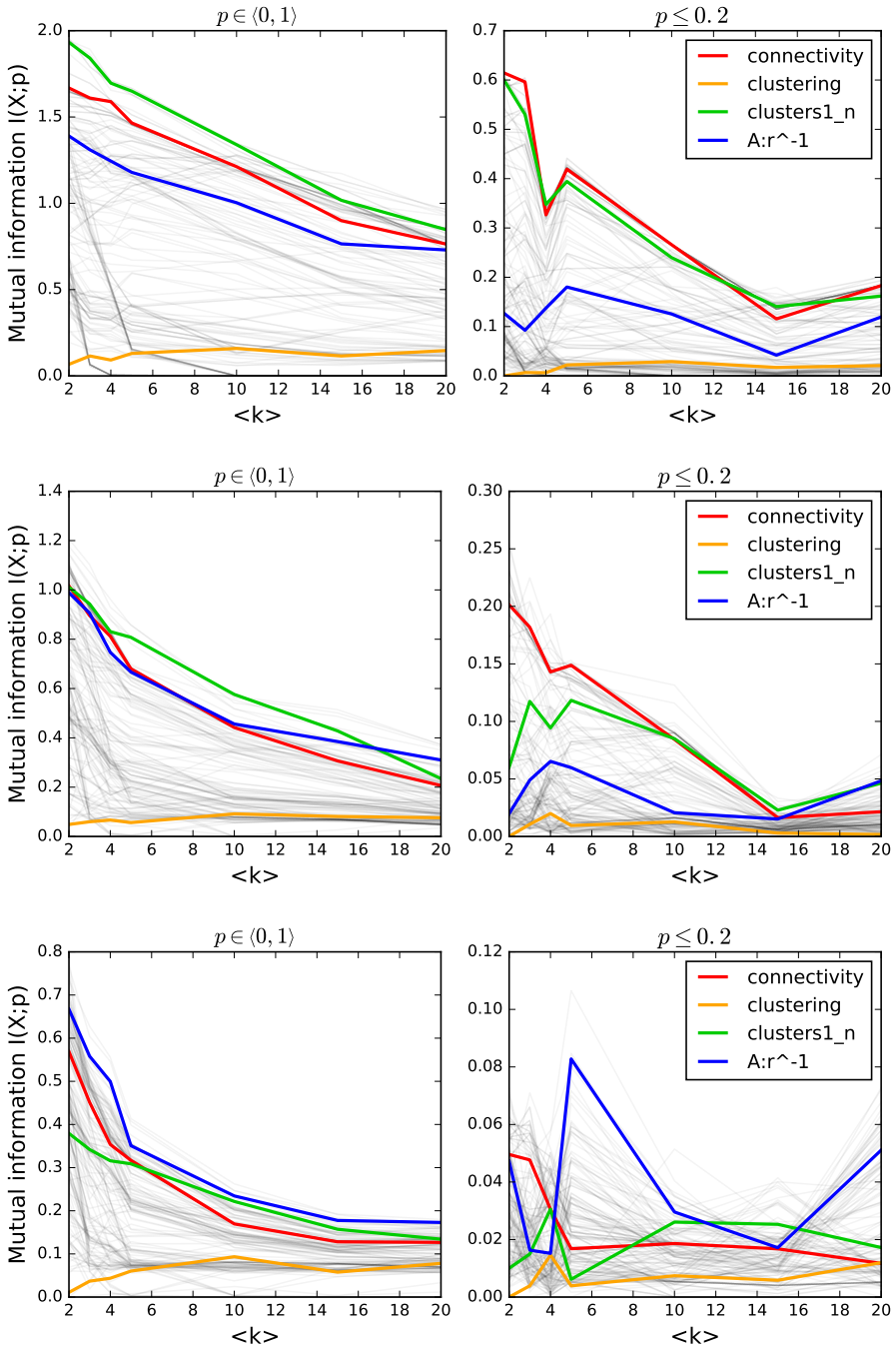


Figure 5: (color online) Same as Figure 3 (ER graphs), but for different visibilities. First row: $q = 1$. Second row: $q = 1/2$. Third row: $q = 1/4$. Note that n has been adjusted according to q, such that the final number of visible colored nodes remains constant (see also Figure S2)

4. Conclusion

We have analyzed the performance of various observables derived from a given subset of nodes (the pattern of 'colored' nodes) to assess, whether this subset is non-randomly distributed in the graph. On a more quantitative level we asked, if and with which accuracy the main parameter of an underlying stochastic process selecting the nodes of this subset (the 'cohesion parameter' p of the 'pattern generator' coloring the nodes) can be inferred from these observables. This basic question has broad implications to any discipline that employs forms of a network-based data interpretation.

Overall, simplicity meets effectiveness as connectivity and edge density are among the best performing observables as regards the inference of the parameter p. Our results show that different types of randomization that could be present during pattern generation lower the inference capabilities in very different ways. For example, observables based on connectivity are much more affected by the environmental range e (i.e., the 'search radius' of the pattern generator) than distance-based measures. The impact of incomplete data, as captured by our visibility parameter q, is of dramatic importance in biological and medical applications [19, 21, 32]. The unexpected robustness to failures in the observation process of distance-based measures is thus a promising result.

Furthermore, the interplay of pattern generator parameters and inference performance of observables also depends on global network topology. Measures based on clustering, for example, perform better in small-world graphs than in scale-free or Erdős-Rényi graphs, whereas for lattices distance-based measures are more efficient, presumably because distances on lattices are much larger than on the other types of networks.

A more comprehensive analysis is required particularly on the following four levels: (1) The vast range of measures, which here only serve as an orientation to put the four highlighted measures in perspective, need to be quantitatively and systematically analyzed. (2) The distinction between weak, intermediate and strong observable signals (corresponding to dense, mildly cohesive and diffuse

patterns, respectively) and the impact of these ranges on inference performance need to be evaluated in more detail. (3) A mechanistic understanding of the impact of the large-scale organization of the graph on the inference of pattern generators needs to be developed. (4) Ultimately, the joint inference of p and the other parameters q and e impacting the observed pattern is to be investigated.

Acknowledgments

M.T. Hütt thanks the Laboratoire de Physique Théorique de la Matière Condensée (LPTMC, Paris) for hospitality and the Physics Institute of the Centre National de la Recherche Scientifique (CNRS, https://www.cnrs.fr/en) for funding his stays, during which part of this work has been performed.

References

[1] D. Stauffer, A. Aharony, Introduction to percolation theory, CRC press, 2018.

[2] D. Stauffer, M. Sahimi, Diffusion in scale-free networks with annealed disorder, Physical Review E 72 (4) (2005) 046128.

[3] M. Rybak, K. Kułakowski, Competing contact processes on homogeneous networks with tunable clusterization, International Journal of Modern Physics C 24 (03) (2013) 1350012.

[4] R. Pandey, D. Stauffer, A. Margolina, J. Zabolitzky, Diffusion on random systems above, below, and at their percolation threshold in two and three dimensions, Journal of Statistical Physics 34 (3-4) (1984) 427–450.

[5] D. Stauffer, A. Aharony, B. Mandelbrot, Self-similarity and covered neighborhoods of fractals: A random walk test, Physica A: Statistical Mechanics and its Applications 196 (1) (1993) 1–5.

[6] P. Erdős, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci 5 (1) (1960) 17–60.

[7] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–2. doi:10.1038/30918.

[8] A. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509.

[9] P. Freche, D. Stauffer, H. Stanley, Surface structure and anisotropy of Eden clusters, Journal of Physics A: Mathematical and General 18 (18) (1985) L1163.

[10] R. Lambiotte, M. Rosvall, Ranking and clustering of nodes in networks with smart teleportation, Physical Review E 85 (5) (2012) 056107.

[11] D. Stauffer, Monte Carlo study of biased diffusion at the percolation threshold, Journal of Physics A: Mathematical and General 18 (10) (1985) 1827.

[12] D. Stauffer, C. Schulze, D. W. Heermann, Superdiffusion in a model for diffusion in a molecularly crowded environment, Journal of Biological Physics 33 (4) (2007) 305.

[13] M.-T. Hütt, Understanding genetic variation – the value of Systems Biology, British Journal of Clinical Pharmacology 77 (4) (2014) 597–605.

[14] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, et al., STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Research 47 (D1) (2019) D607–D613.

[15] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, et al., The BioGRID interaction database: 2019 update, Nucleic Acids Research 47 (D1) (2019) D529–D541.

[16] I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, et al.,

A community-driven global reconstruction of human metabolism, Nature Biotechnology 31 (5) (2013) 419–425.

[17] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G. A. P. Gonzalez, M. K. Aurich, et al., Recon3D enables a three-dimensional view of gene variation in human metabolism, Nature Biotechnology 36 (3) (2018) 272.

[18] L. Licata, P. Lo Surdo, M. Iannuccelli, A. Palma, E. Micarelli, L. Perfetto, D. Peluso, A. Calderone, L. Castagnoli, G. Cesareni, SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update, Nucleic Acids Research 48 (D1) (2020) D504–D510.

[19] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabási, Uncovering disease-disease relationships through the incomplete interactome, Science 347 (6224).

[20] J. K. Huang, D. E. Carlin, M. K. Yu, W. Zhang, J. F. Kreisberg, P. Tamayo, T. Ideker, Systematic evaluation of molecular networks for discovery of disease genes, Cell Systems 6 (4) (2018) 484–495.

[21] L. Cowen, T. Ideker, B. J. Raphael, R. Sharan, Network propagation: a universal amplifier of genetic associations, Nature Reviews Genetics 18 (9) (2017) 551.

[22] N. Sonnenschein, M. Geertz, G. Muskhelishvili, M.-T. Hütt, Analog regulation of metabolic demand, BMC Systems Biology 5 (1) (2011) 40.

[23] N. Sonnenschein, J. F. G. Dzib, A. Lesne, S. Eilebrecht, S. Boulkroun, M.-C. Zennaro, A. Benecke, M.-T. Hütt, A network perspective on metabolic inconsistency, BMC Systems Biology 6 (1) (2012) 41.

[24] C. Knecht, C. Fretter, P. Rosenstiel, M. Krawczak, M.-T. Hütt, Distinct metabolic network states manifest in the gene expression profiles of pediatric inflammatory bowel disease patients and controls, Scientific Reports 6 (2016) 32584.

[25] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science 342 (6164) (2013) 1337–1342.

[26] A. Messé, M.-T. Hütt, P. König, C. C. Hilgetag, A closer look at the apparent correlation of structural and functional connectivity in excitable neural networks, Scientific Reports 5 (2015) 7870.

[27] A. Messé, M.-T. Hütt, C. C. Hilgetag, Toward a theory of coactivation patterns in excitable neural networks, PLoS Computational Biology 14 (4) (2018) e1006084.

[28] L. Turnbull, M.-T. Hütt, A. A. Ioannides, S. Kininmonth, R. Poeppl, K. Tockner, L. J. Bracken, S. Keesstra, L. Liu, R. Masselink, et al., Connectivity and complex systems: learning from a multi-disciplinary perspective, Applied Network Science 3 (1) (2018) 11.

[29] D. Stauffer, Monte Carlo study of density profile, radius, and perimeter for percolation clusters and lattice animals, Physical Review Letters 41 (20) (1978) 1333.

[30] P. Nyczka, M.-T. Hütt, Generative network model of transcriptome patterns in disease cohorts with tunable signal strength, Phys. Rev. Research 2 (2020) 033130. doi:10.1103/PhysRevResearch.2.033130.
URL https://link.aps.org/doi/10.1103/PhysRevResearch.2.033130

[31] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Physical Review E 69 (6) (2004) 066138.

[32] J. Sanz, E. Cozzo, J. Borge-Holthoefer, Y. Moreno, Topological effects of data incompleteness of gene regulatory networks, BMC Systems Biology 6 (1) (2012) 110.

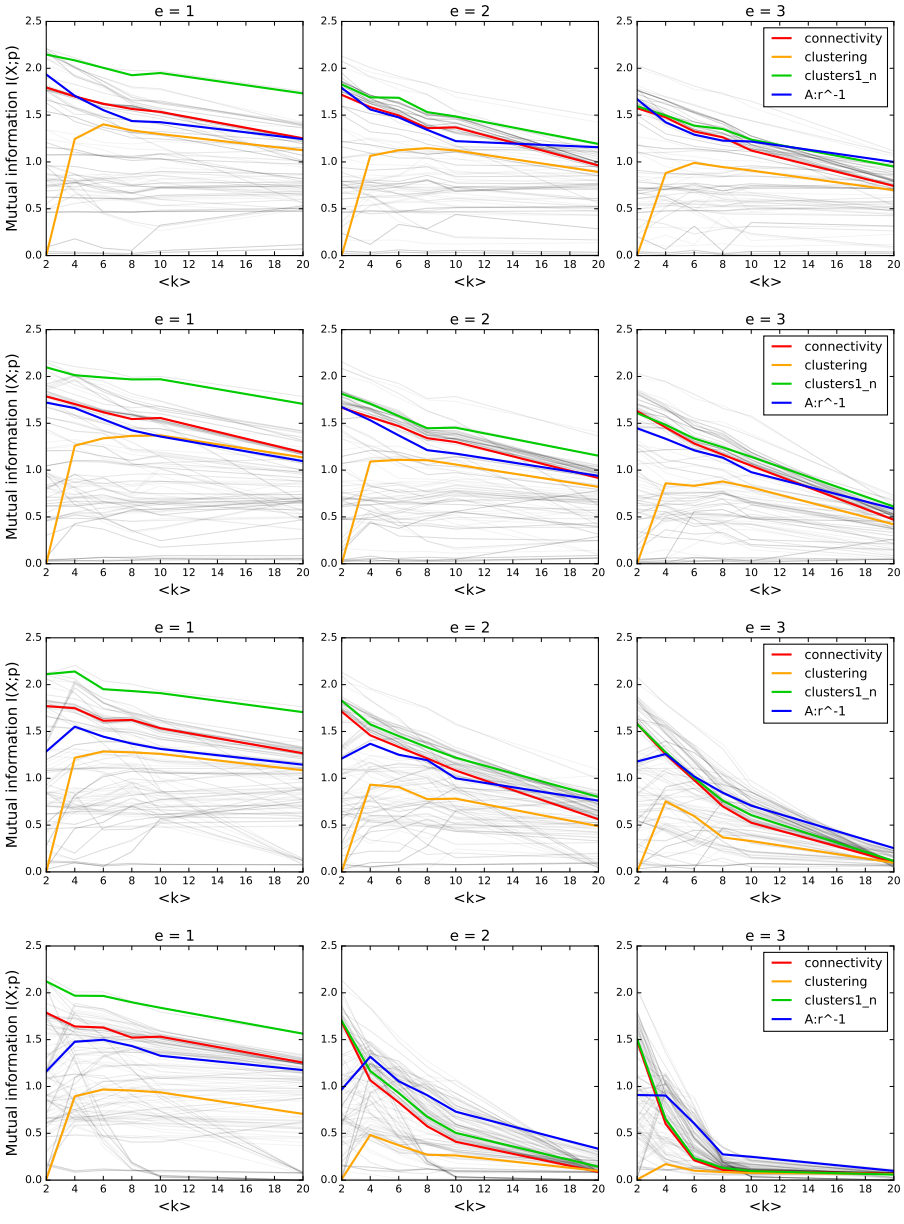


Figure S1: Same as the lower panel (b) of Figure 4, but only for small-world graphs (SW) with different values of network parameter m of the Watts-Strogatz model. Left column: $e = 1$. Middle column: $e = 2$. Right column: $e = 3$. First row: $m = 0$. Second row: $m = 0.001$. Third row: $m = 0.01$. Fourth row: $m = 0.1$.

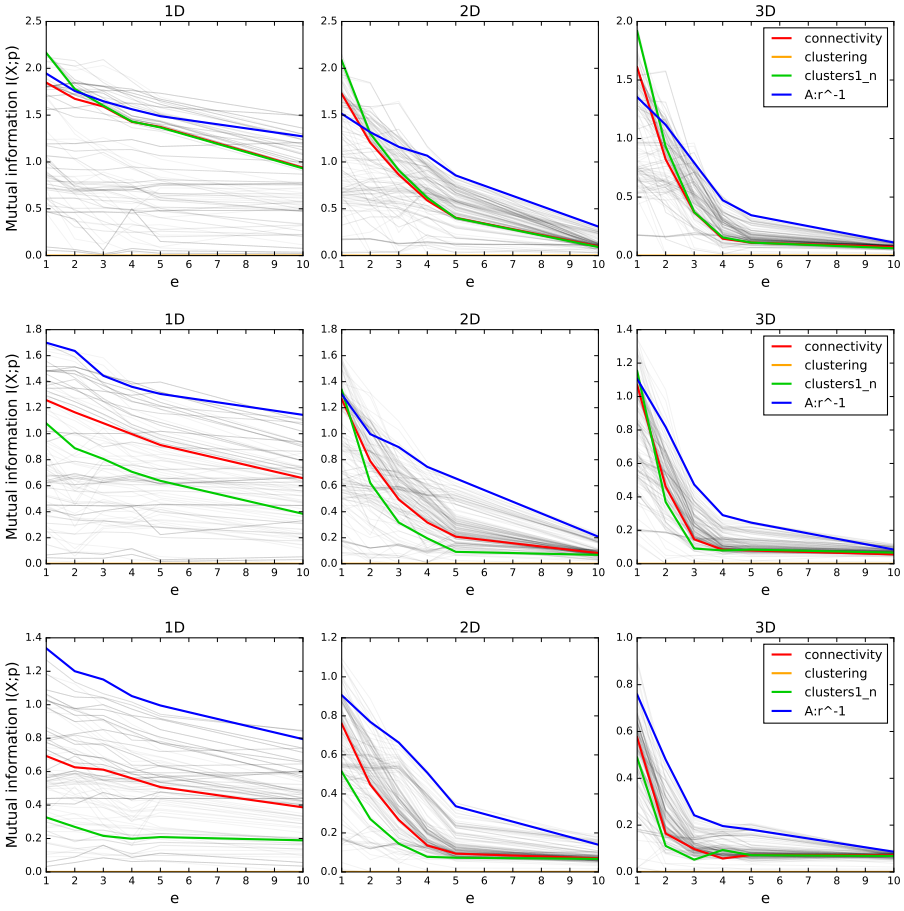


Figure S2: Each row presents a set of three lattices of increasing space dimension: 1D, 2D and 3D. Different rows present different values of q and n, where n is tuned to the value of q in order to keep $\langle n' \rangle = const. = 50$. First row: $n = 50$, $q = 1.0$. Second row: $n = 100$, $q = 0.5$. Third row: $n = 200$, $q = 0.25$.