Using symbolic networks to analyse dynamical properties of disease outbreaks

José L. Herrera-Diestra¹,², Javier M. Buldú³,⁴,⁵, Mario Chavez⁶ and Johann H. Martínez¹,³,⁷

¹ICTP South American Institute for Fundamental Research, Instituto de Física Teórica, Universidade Estadual Paulista, São Paulo, Brazil
²CeSiMo, Facultad de Ingeniería, Universidad de Los Andes, Mérida, Venezuela
³Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain
⁴Laboratory of Biological Networks, Center for Biomedical Technology - UPM, Madrid, Spain
⁵Complex Systems Group, Universidad Rey Juan Carlos, Móstoles, Spain
⁶CNRS UMR7225, Hôpital Pitié Salpêtrière, Paris, France
⁷Department of Biomedical Engineering, Universidad de los Andes, Bogotá, Colombia

We introduce a new methodology, which is based on the construction of epidemic networks, to analyse the evolution of epidemic time series. First, we translate the time series into ordinal patterns containing information about local fluctuations in disease prevalence. Each pattern is associated with a node of a network, whose (directed) connections arise from consecutive appearances in the series. The analysis of the network structure and the role of each pattern allows them to be classified according to the enhancement of entropy/complexity along the series, giving a different point of view about the evolution of a given disease.

1. Introduction

Time-series analysis (TSA) is a successful field with a cross-disciplinary character. In economics, climate science, geophysics and statistics [1–4], among others, forecasting based on observation of the previous states of a system is of critical interest. Engineering and communications are focused on the control and parameter detection of signals [5–7]. Classification,
clustering or detection of abnormalities in a collection of samples are the main targets for data mining and machine learning [8,9]. A time series is basically the collection of a given variable describing the activity of a system during different time points. This system may be linear or nonlinear; nevertheless, it translates this property to the observed time series. Regardless of the diversity of techniques used to acquire and study a collection of temporal samples, the extraction of information about the underlying dynamical system by observing the evolution of one (or several) of its variables is frequently a non-trivial task [10,11].

Notwithstanding the difficulties, several methodologies and tools have been proposed to better capture and understand the properties of a system from the observation of its dynamics. In general, they rely on the fact that time series have internal structures along several dimensions (time, amplitude, phase, frequency, etc.) that contain information about the dynamical response of a system. The way these structures are related helps to describe the system’s activity and to predict its evolution. In this regard, TSA has broadly focused on two perspectives. The first perspective consists of (i) a situation in which a few parameters can describe time series if they come from stationary stochastic processes and (ii) a non-parametric situation in which the estimation of the spectral density or higher-order conditional moments totally describes a collection of samples [12,13]. In the second perspective, TSA methods can be grouped according to whether they are univariate or multivariate. The latter accounts for techniques that quantify the contribution of two or more variables to a single event, while the former aims at describing and inferring the evolution of temporal structures based on a single variable [14–17]. Regarding univariate methods, a central question is to know whether temporal structures correlate with each other. Serial correlation quantifies the point-by-point correlation of a signal with a delayed version of itself [18]. Autocorrelation is then useful to accurately seek repetitive patterns in a signal. However, when a signal is split into consecutive segments, autocorrelation fails to capture the interplay between them. Therefore, a naive but not trivial question may raise about whether signal segments of a time series might be related to each other.

In this context, this paper is focused on the hypothesis that consecutive segments of a time series may be inter- related, transferring information along them. To test this hypothesis, we unravel how segments transmit information between them, with the final objective of gaining additional insights into the properties of time series and, ultimately, into the system behind them. We quantify the communication levels among consecutive segments of a time series using a combination of network science (NS) and symbolic dynamics (SD).

The use of NS has reached a broad range of areas, such as social sciences, biology, ecology, neuroscience and epidemics, among others [19]. Its main advantage is its ability to translate almost any system under interaction into nodes, which are endowed with properties that can be extracted from almost any variable of the system. These nodes are connected by links, which are entities that account for any kind of interaction between them. The mathematical background of NS is useful when designing artificial models that lead to different network structures, while its transversal nature allows these models to be applied to different kinds of datasets [19,20]. NS has also been applied in data mining and machine learning problems, and more recently it has been implemented as a concomitant tool of TSA [21–28].

In this context, publications using visibility graphs have introduced new perspectives on how to extract information from experimental time series [26,29]. The idea behind visibility graphs is that a time series can be transformed into a network, where nodes are the different values of the time series and their links are created between any two amplitudes as long as they can ‘see’ each other without being covered by another intermediate sample, as if it were a series of concatenated hills and valleys [26]. Since the seminal work of Lozano-Perez & Wesley [29], visibility graphs have been applied to a broad range of problems, from planning collision-free paths to avoid polyhedral obstacles, characterization of random time series and combinatorics on words to three-dimensional perspectives for image processing [30].

Alternative techniques to map time-series dynamics into a graph representation have been introduced by means of symbolic dynamics [31–35]. SD assumes a symbol as a perceptible idea that encompasses the common characteristics of a set of samples in a time series. In other
words, a symbol contains valuable information about a temporal segment but, at the same time, it simplifies the analysis of the original variable. The symbol definition has led to different approaches, many of them with interesting results [36–39]. Among the diversity of definitions of a symbol, the one proposed by Bandt & Pompe [40] is nowadays the definition that is most endorsed by the community of researchers making use of symbolic time series. The originality of this approach consists of a formalism that quantifies the information of a time series by its transformation into ordinal patterns, which take into account the ranking between consecutive values.

An ordinal pattern $\pi$ is defined by comparing the relative amplitudes of a temporal segment of $D$ consecutive element observations ($D$ is also called the dimension of the pattern), and mapping them into a one-dimensional ordinal space \{0, 1, $\ldots$, $(D - 1)$\}, where the highest value in the segment will be tagged by $D - 1$ and the lowest will be assigned 0, with the rest of the values going in descending order. Thus, each element of the pattern $\pi$ only contains the order, disregarding its specific value. For finite time series of length $M$, when the cardinality $D$ of the temporal segments is fixed, the number of possible symbols is determined by $D!$. In this way, the reduction in the variables boosts the run-time of the exploratory experiments. Next, the probability distribution of finding a pattern $\pi$ in the symbol sequence is obtained and analysed in order to obtain information about the general properties of the underlying system. This methodology has been proved to be robust against noise and is fully data-driven, adequate under weak stationarity and computationally efficient with no further assumptions over the datasets [39,41–46]. It has been extensively used to characterize time series of different natures, as well as for mapping a collection of samples into a directed graph. In this regard, a directed graph containing $D!$ nodes can be created when different symbols $\pi$ of a signal are consecutively connected as they appear in the time series [28]. Using this transformation of patterns into networks, lasers and chaotic time series were characterized by means of the link entropy and a non-normalized version of the Shannon entropy of the resulting nodes. This work revealed the usefulness of mapping time series into a graph, and opened the door to further studies about the structure of the resulting networks.

In this paper, we go one step beyond by studying the transfer of information along temporal segments of epidemic time series. We transform them into ordinal patterns, which are then projected into nodes of a network whose links are based on consecutive appearances along the time series. In particular, we construct symbolic networks from epidemic time series ($x_t$) of vector-borne (dengue fever, malaria) and air-borne (influenza) diseases reported in different countries (see §2). We also introduce a family of five novel parameters that characterize the role of the nodes (patterns) of the network, specifically: (i) the amount of entropy entering/leaving a node $\pi$, by means of the incoming/outgoing entropy ($H_{in}, H_{out}$), (ii) the amount of complexity entering and leaving a node $\pi$, by means of the incoming/outgoing complexity ($C_{in}, C_{out}$), (iii) the level of conductance of information of a node, by means of the flux of entropy/complexity ($\phi_{H,C}$), (iv) how much a node amplifies or attenuates entropy/complexity ($A_{H,C}$), and (v) the internal fluctuation ($f$) inside each pattern. Using these five metrics, we are able to identify differences between diseases and assign specific roles to the different patterns of the time series. We explore the interplay between the fluctuation $f$ of each pattern $\pi$ and the role the corresponding node has on the network structure. Next, we compare our results with a set of synthetic outputs of different complexity in order to infer the closeness similarities of these diseases with synthetic models. Our results demonstrate that this methodology unveils differences among air-borne and vector-borne diseases, and evidences how these outbreaks share similarities with autoregressive processes of a linear and nonlinear nature.

2. Methods

(a) Datasets

The epidemiological cohorts are composed of nine time series $x_t$ of different length $M$ corresponding to three diseases in six different countries for the time periods shown in table 1.
the entropy and complexity (vectors (nodes), leading to a weighted graph. Next, we analyse networks of diseases using a sequentially more than one time, a weight is given to the direct connection between the two relative amplitudes of each vector into an ordinal representation. When two vectors co-occur then consecutively connected in order of appearance, which leads to a directed graph. Patterns associated with each node are extracted from a symbolic transformation mapping the inner relative amplitudes of each vector into an ordinal representation. When two vectors co-occur sequentially more than one time, a weight is given to the direct connection between the two vectors (nodes), leading to a weighted graph. Next, we analyse networks of diseases using a family of novel network parameters, consisting of a set of information-based metrics quantifying the entropy and complexity (in/out entropy, in/out complexity) of the nodes and a set of parameters assessing their dynamical role (flux, amplitude and fluctuation).

(b) Symbolic transformation

Following the methodology proposed by Bandt & Pompe [40], we retrieve all different patterns \( \pi \) of length \( D \) appearing in a signal \( x_t \). When defining an embedding dimension \( D \), we are constrained by the size \( D \) of the patterns, since \( D! \) is the total number of possible patterns, e.g. \( D = 3 \) leads to \( D! = 6 \) possible symbols emerging from \( x_t \). The number of all appearing symbols is also closely related to the length \( M \) of the time series, since extremely short time series do not have enough statistics to guarantee a fully resolved distribution of patterns. For a statistically reliable estimation, we follow the condition \( M - D \gg D! \) [53]. Once this condition is fulfilled, each symbol is then constructed by considering consecutive samples of length \( D \) extracted from \( x_t \). Under this mapping, \( x_t (\forall t = 1, 2, \ldots, M) \) is transformed into a restricted number of patterns encoding the relative inner amplitudes of the \( D \)-dimensional vector \( \{x_t, x_{t+1}, \ldots, x_{t+D} \} \). Samples are arranged (or ranked) onto the permutation \( \pi = (\pi_0, \pi_1, \ldots, \pi_{D-1}) \) of \( (0, 1, \ldots, D - 1) \) fulfilling \( x_{t+\pi_0} \leq x_{t+\pi_1} \leq x_{t+\pi_{D-1}} \). Hence, each permutation is a symbol of the full spectrum of available patterns. The full process is summarized, schematically, in figure 1. This methodology has been proven to be computationally efficient, fully data-driven with no further assumptions on the data, robust against noise and well-behaved under weak stationarity [42–46,54–57].

(i) Pattern fluctuation

We propose to quantify the variability of each symbol by defining its fluctuation (f). Let us consider pattern \( \pi_t \) placed in what we call the order plane (figure 1d). For each pattern, we calculate the sum of the distances between each point of the pattern, labelled as \( T_i \). After calculating \( T_i \), we sort all possible values in increasing order and group them. Patterns with the lowest \( T_i \) are assigned the rank \( f = 1 \), the next value of \( T_i \) (all patterns with this value) are

| country   | dengue     | influenza  | malaria   |
|-----------|------------|------------|-----------|
| Australia | 974 (1997–2015) | 626 (2005–2016) | 669 (2002–2014) |
| Colombia  | 626 (2005–2016) | 626 (2005–2016) | 830 (2000–2015) |
| Japan     | 964 (1998–2015) | 964 (1998–2015) | 830 (2000–2015) |
| Mexico    | 678 (2000–2015) | 830 (2000–2015) | 830 (2000–2015) |
| Singapore | 838 (2000–2015) | 830 (2000–2015) | 830 (2000–2015) |
| Venezuela | 660 (2002–2014) | 669 (2002–2014) | 669 (2002–2014) |

Each time series represents the weekly number of individuals infected with a specific disease. Before applying our method, we applied the variance-stabilizing transformation \( y_t = \log(x_t + 1) \).

The datasets analysed in this paper were extracted from online reports of the corresponding Ministry of Health of each country [47–52]. We studied diseases in terms of the graph representation of their aggregated time series. Time series are cut into vectors of specific length \( D \), where each different vector corresponds to a node of the network. Nodes (vectors) are then consecutively connected in order of appearance, which leads to a directed graph. Patterns associated with each node are extracted from a symbolic transformation mapping the inner relative amplitudes of each vector into an ordinal representation. When two vectors co-occur sequentially more than one time, a weight is given to the direct connection between the two vectors (nodes), leading to a weighted graph. Next, we analyse networks of diseases using a family of novel network parameters, consisting of a set of information-based metrics quantifying the entropy and complexity (in/out entropy, in/out complexity) of the nodes and a set of parameters assessing their dynamical role (flux, amplitude and fluctuation).
Figure 1. Extracting ordinal patterns from a time series. (a) Signal amplitudes are split into consecutive vectors of length $D$ (in this example, $D = 3$). (b) The values of the amplitude inside each vector are translated into a ranking. Consecutive patterns are connected through a direct link. (c) Directed graph derived from the time series and patterns. (d) For illustrative purposes, vector $[-0.3, 1.2, 0.5]$ is transformed into the ordinal pattern $\pi = \{0, 2, 1\}$ and its variability is obtained as $T_i = d_1 + d_2 = \sqrt{5} + \sqrt{2} = 3.65$ (where $d_1$ and $d_2$ are the lengths of the lines connecting two consecutive rankings), corresponding to a fluctuation of $f = 2$ (see Table 2 and main text, for details). (Online version in colour.)

Table 2. Pattern fluctuations for $D = 3$. First row: all possible patterns $\pi$ of length 3. Second row: the internal variability $T$ (obtained as the length of the connecting lines). Third row: the corresponding fluctuation parameter $f$. Note that different patterns have the same fluctuation, since they have the same internal variability.

| $\pi$ | $T$  | $f$ |
|-------|------|-----|
|       | 2.828| 1   |
|       | 2.828| 1   |
|       | 3.650| 2   |
|       | 3.650| 2   |
|       | 3.650| 2   |
|       | 3.650| 2   |

assigned the rank $f = 2$, and so on until all values of $T_i$ are covered (see Table 2 for details). In this way, we can group patterns in terms of their variability, which gives interesting information about the temporal dynamics of the patterns. For instance, monotonic and periodic time series lead to fluctuations with low $f$ (decreasing/increasing patterns). We believe that $f$ is a parameter that complements the information about the specific features of a given node.

(ii) Network construction

We built the network representation of the time series by transforming each different pattern into a node, following the methodology proposed by Masoller et al. [28]. Consecutive symbols were connected through a direct link (see red arrows in figure 1b) and auto-loops (links leaving and reaching the same node) were dismissed. In this way, sequential associations of patterns lead to a
directed graph, while repetitions of pattern co-occurrence (e.g., when pattern \( \pi_i \) and \( \pi_j \) appear one after the other more than once) reinforce the weight \( w_{ij} \) of the link going from node \( i \) to \( j \) (in other words, from pattern \( \pi_i \) to \( \pi_j \)). The weight of a link \( \pi_i \rightarrow \pi_j, w_{ij} \), is the relative number of times, in the sequence, the symbol \( i \) is followed by symbol \( j \); in this way, the link weights are normalized in each node, i.e. \( \sum_{j=1}^{D_i} w_{ij} = 1 \). Figure 1c shows the resulting network from the time series and patterns.

(c) Global measures

Entropy \( H \) and statistical complexity \( C \) are useful when characterizing the dynamics of several systems ([58–61] and references therein), as well as when characterizing the appearance of ordinal pattern populations along the time series [46,57]. A simple diagram of these measures globally characterizes the time series dynamics (figure 2). In this way, a region of positive correlations between \( H \) and \( C \), which is highlighted in blue, while the region of negative correlation is highlighted in yellow. (Online version in colour.)

Figure 2. Qualitative representation of the entropy–complexity \((H \times C)\) plane. Entropy \( H \) increases with the disorder of the signal. Disequilibrium \( Q \) decays as the signal approaches a complete disordered state. Statistical complexity \( C = Q \cdot H \) reaches its maximum when the system stays between order and disorder. Now the region of positive correlations between \( H \) and \( C \), which is highlighted in blue, while the region of negative correlation is highlighted in yellow. (Online version in colour.)

(d) Local measures

In addition to the pattern fluctuation \( f \), we also propose a collection of node features based on global parameters. We define the incoming content of information of a node \( i \) as the probability distribution \( p_{in} = \{p_l\} \) of the inbound links, \( \forall l = \{1, \ldots, k_{in}\} \), where \( k_{in} \) is the in-degree. To do that, the weight of each incoming link is divided by the total weight of the \( l \) links entering \( i \), such
that $\sum_{i=1}^{k_i} p_i = 1$. Reciprocally, we construct the outgoing information content $p_i^{\text{out}} = \{p_m\}$ upon outgoing $m$ edges from $i \forall m = \{1, \ldots, k_{\text{out}}\}$, with $k_{\text{out}}$ as the out-degree. All departing weights are divided by the total sum of its $m$ outbound weights fulfilling $\sum_{m=1}^{k_i} p_m = 1$. This turns a node $i$ into an entity endowed with the property of conveying information among ordinal patterns, which can be interpreted as the way temporal structures (patterns) communicate (or transmit information) along time.

(i) Nodal entropy and complexity

Under this framework, we can measure the Shannon entropy arriving at node $i$ as $S_i = -\sum_{i=1}^{k_i} p_i \ln(p_i)$, and we can contrast it with its maximal entropy $S_{i,\text{max}} = \log(k_i)$ obtained for the uniform distribution $p_i = (p_{i,e} | p_{i,e} = 1/k_i \forall i \forall l)$. The ratio between $S_i$ and $S_{i,\text{max}}$ leads to the incoming node entropy $H_{\text{in}}[p_i] = S_i / S_{i,\text{max}}$, which is bounded by $0 \leq H_{\text{in}}[p_i] \leq 1$. When only one link arrives at node $i$, i.e. $p_i^0 = 1$, the incoming node entropy is at its lowest ($H_{\text{in}} = 0$). On the other hand, when all possible links arrive at $i$ (i.e. node $i$ has $l = D! - 1$ incoming links), the incoming node entropy is at its highest ($H_{\text{out}} = 1$) if all incoming links have the same probability $p_i = 1/k_i$. Analogous reasoning applied to the $m$ outgoing links leads to the outgoing node entropy $H_{\text{out}}[p_m]$. The first row of table 3 summarizes the definitions of the incoming and outgoing node entropies.

Similar to global measures, by comparing with a uniform distribution $p_e$, one can measure the disequilibrium of incoming links $Q_{\text{in}}[p_i] = Q_0 \cdot D[p_i, p_e]$. Here, $Q_0 = -2((l_{\text{out}} + 1)/k_{\text{in}}) \log(k_{\text{in}} + 1) - 2 \log(2k_{\text{in}}) + \log(k_{\text{in}}))^{-1}$ is the normalization constant leading to $0 \leq Q_{\text{in}}[p_i] \leq 1$. $D[p_i, p_e]$ accounts for the Jensen–Shannon divergence defined in terms of $S_i$ as $D[p_i, p_e] = S_i[p_i + p_e]/2 - S_i[p_i]/2 - S_i[p_e]/2$.

By multiplying $H_{\text{in}}$ and $Q_{\text{in}}$, we finally obtain the incoming node complexity that is bounded between $0 \leq C_{\text{in}}[p_i] \leq 1$. Following the same procedure as node entropy, we can define the outgoing node complexity $C_{\text{out}}[p_m]$ considering the $m$ outgoing links. The second row of table 3 summarizes the definitions of the incoming and outgoing node complexities.

The previous parameters shed light on how temporal structures in time series are temporally correlated with each other, and how different patterns are related to the increase in entropy and complexity along a time series. Additionally, we introduce two new measures to characterize how nodes process the entropy/complexity that passes through them.

(ii) Flux and amplification of information

The flux is a vectorial quantity that describes the magnitude and direction of the flow of information passing through a node $i$. In directed graphs, direction is trivially associated with incoming links of node $i$ to its outgoing ones. Therefore, the net amount of ‘information’ passing through $i$ can be quantified in terms of two nodal properties: the entropy flux $\phi_H$ and complexity flux $\phi_C$ fluxes. Fluxes $\phi_H$ and $\phi_C$ are measures of the dynamical importance of a node, since they characterize which patterns behave as dynamical hubs by allowing larger amounts of information to enter and leave the pattern $i$. They quantify the dynamical relevance of nodes by taking into account the entropies and complexities of the incoming and outgoing links. Table 4 contains the mathematical definitions of the entropy and complexity fluxes.

Importantly, fluxes allow the identification of basin nodes, which retain all incoming information from other patterns and do not transmit information to others, i.e. nodes whose entropy (or complexity) flux is $\phi_H = \sqrt{H_{\text{in}}^2 + H_{\text{out}}^2} = H_{\text{in}}$. On the other hand, generator nodes have no incoming links, but convey information to other nodes $\phi_H = \sqrt{H_{\text{out}}^2 + H_{\text{out}}^2} = H_{\text{out}}$. However,

| Table 3. Definitions of the node entropies based on the distribution of incoming and outgoing links of node $i$. |
|----------------------------------|
| nological entropy $H_{\text{in}}[p_i] = S_i / S_{i,\text{max}}$ | $H_{\text{out}}[p_m] = S_m / S_{m,\text{max}}$ |
| nological complexity $C_{\text{in}}[p_i] = H_{\text{in}}[p_i] \cdot Q_{\text{in}}[p_i]$ | $C_{\text{out}}[p_m] = H_{\text{out}}[p_m] \cdot Q_{\text{out}}[p_m]$ |
Table 4. Definitions of the flux $\phi$ and amplification $A$ of a node $i$, both for the entropy and complexity of the incoming and outgoing links.

| Type          | Amplifier | Transmitter | Attenuator | Basin |
|---------------|-----------|-------------|------------|-------|
| $\phi_{H,C}$  | $H_{out}$ | $H_{out}$   | $H_{in}$   | $H_{in}$ |
| $\phi_{C}$    | $C_{out}$ | $C_{out}$   | $C_{in}$   | $C_{in}$ |
| $A_{H,C}$     | n.a.      | $> 1$       | $1$        | $< 1$  |
| $A_{C}$       | $0$       | $0$         | $0$        | $0$    |

Table 5. Definitions of the dynamical roles of a node. The first row depicts five types of transitions for a pattern $\pi_i$, $P_i$, $P_m$ are the probabilities associated with incoming and outgoing links, respectively. The second row corresponds to the role assigned to the nodes. The third row contains the limits and domains for the flux $\phi_{H,C}$. Note that $\phi_{H,C} = \sqrt{2}$ only for dynamical hubs. The last row shows the boundaries of $A_{H,C}$. Here n.a. stands for not applicable.

![Diagram]

note that generator nodes may not exist in our kind of networks, since all patterns appear after a previous one (with the unique exception of the first pattern in the time series). Finally the flux dynamical hubs are those patterns receiving large amounts of information from other temporal structures (i.e. patterns that are reached from a large number of different patterns), and redistribute it equally (i.e. also lead to a diversity of patterns). For example, in the extreme case of $(\phi_H = \sqrt{1^2 + 1^2} = \sqrt{2})$ we would have the most important hub for the entropy flux. As a consequence, by definition, the values of fluxes are bounded by $0 < \phi_{H,C} \leq \sqrt{2}$.

The last dynamical feature we introduce is the entropy $A_H$ and complexity $A_C$ amplifications, whose definitions are given by equations (2.2b) in table 4. While $\phi_{H,C}$ give us an idea of which patterns act as pipe flows, $A_{H,C}$ tell us about the node gain. In other words, the ability to amplify or diminish the level of information a pattern is transmitting. In the case of entropy, for instance, basin nodes have zero amplification $A_H = 0 / H_{in}$. Transmitter nodes that equally receive and distribute information would reach up to $A_H = 1$ when $H_{out} = H_{in}$, as in the case of dynamical hubs. Nodes that increase levels of entropy will have $A_H > 1$ and are called amplifier nodes. In the case of $A_H$, they amplify the level of entropy received from the incoming links, taking the system to a disordered one. In the time series, they correspond to short periodic temporal structures followed by irregular patterns. In the case of complexity, an amplifier node receives links endowed with either lower or higher entropy, but in both cases outgoing links have higher complexity levels. In the same way, a node might receive higher levels of entropy, taking the system down to lower ones. These attenuator nodes ($A_H < 1$) can be related to changes from irregular fluctuations to more periodic patterns. Likewise, in terms of $A_C$, attenuator nodes reduce the complexity levels of the system.

Table 5 summarizes the node classification based on incoming and departing probabilities $P_i$ and $P_m$. Each dynamical role is characterized by the two fluxes $\phi_{H,C}$ and the two amplifications $A_{H,C}$.

3. Results

Networks obtained from the same disease are grouped to visualize, first, the incoming and outgoing complexity of a node $i$; respectively, $C_{in}$ and $C_{out}$ (see equation (2.1b) in Table 3 for
Figure 3. Complexity entering a node (pattern) $C_{\text{in}}$ versus complexity leaving a node $C_{\text{out}}$. Colour codes are the same throughout the paper: red for dengue (a), blue for influenza (b) and green for malaria (c). Dashed lines correspond to $C_{\text{out}} = C_{\text{in}}$. Solid lines are the regression lines. (a) Dengue coefficient of determination is $R^2 = 0.6$. (b) Influenza has $R^2 = 0.71$. (c) In Malaria, $R^2 = 0.3$. (Online version in colour.)

As shown in figure 3, we found positive correlations between these two variables in all diseases, which means that the higher the complexity entering a node, the higher the complexity departing from it (note that dashed lines correspond to $C_{\text{in}} = C_{\text{out}}$). Influenza (figure 3(b)), which is an airborne disease, reaches higher levels of complexity, with nine nodes with complexities higher than 0.10 (only two in the case of dengue and none in malaria). By contrast, malaria networks (figure 3c) contain nodes with the lowest amounts of complexity.

We have obtained linear regressions accounting for the interplay between $C_{\text{out}}$ and $C_{\text{in}}$ for the three diseases. Influenza has the highest coefficient of determination $R^2 = 0.71$, which means that a linear equation is around 70% effective at estimating the values of $C_{\text{out}}$. Vector-borne diseases have lower $R^2$, in comparison with influenza. Interestingly, malaria has the lowest bound with a value of $R^2 = 0.3$, which reveals the absence of a linear correlation between $C_{\text{out}}$ and $C_{\text{in}}$. Since the dashed lines in figure 3 correspond to $C_{\text{in}}=C_{\text{out}}$, nodes (i.e. patterns) lying above these lines are complexity amplifiers (see the classification of node roles in table 5), indicating that they distribute more complexity than the levels they receive. On the contrary, nodes below the dashed lines act as complexity attenuators, since they reduce the complexity existing in previous patterns.

In order to understand how the complexity and entropy of a disease evolve, and what are the patterns that increase or decrease them, we obtain the flux $\phi$ and the amplification $A$ for both $C$ and $H$ (see Methods for details). Figure 4 shows the interplay between the entropy and complexity fluxes, respectively $\phi_H$ and $\phi_C$, that a node handles.

Figure 4 shows the behaviour of the three diseases, revealing negative correlations between entropy and complexity fluxes. This result indicates that we are in a region where the level of stochasticity is high, since, as we can see in figure 2, entropy and complexity only have positive correlations in situations of high disorder.

Also note that entropy hubs, i.e. those nodes handling high amounts of entropy, are located on the bottom right-hand side of figure 4a–c. These nodes receive the highest amount of entropy and distribute it among the rest of the patterns in the network, but fail as complexity distributors.
Interestingly, influenza and malaria have two entropy hubs that reach the highest possible value ($\phi_H = \sqrt{2}$) and, as a consequence, their complexity flux decreases to $\phi_C = 0$ (see Methods for an explanation of the boundaries of $\phi_H$ and $\phi_C$).

Similar to figure 3, patterns in influenza span their $\phi_C$ along different levels of $\phi_H$, reaching the highest complexity values. This suggests that some specific patterns in influenza behave as units of good ‘conductance’ of complexity, while others act as entropy transmitters. On the contrary, malaria only has entropy hubs, while dengue seems to be between the other two diseases, having just one complexity hub with $\phi_C > 1.2$. Linear regression of the interplay between $\phi_C$ and $\phi_H$ shows high values of the coefficient of determination $R^2$ for the three diseases.

While entropy and complexity fluxes $\phi_{H,C}$ allow the existence of hub patterns and their conductance level to be determined, the amplification $A$ gives an estimate of how much entropy/complexity a node gains or loses in the network.

Amplification is defined as the ratio between the incoming and outgoing entropy/complexity that passes through a node $i$ (see equation (2.2b) in table 4 for the mathematical definition). In figure 5, we plot the $(A_H, A_C)$ plane of the three diseases, where we can observe that the amplification parameter allows us to define four regions of interest (marked by dashed lines) that, in turn, assign different roles to the nodes of the disease network. Region $R_1$ includes nodes that amplify both entropy ($A_H > 1$) and complexity ($A_C > 1$). Since, as we have previously seen, we are in a state where a negative correlation exists between entropy and complexity, there are no nodes lying within this region. Region $R_2$ defines nodes that may act as entropy attenuators ($A_H < 1$) and complexity amplifiers ($A_C > 1$). Region $R_3$ allocates nodes that attenuate both entropy ($A_H < 1$) and complexity ($A_C < 1$). Finally, region $R_4$ corresponds to those nodes that increase the entropy ($A_H > 1$) while reducing their complexity ($A_C < 1$).

The case of $A_H = A_C = 1$ (i.e. no amplification is reported) only occurs for nodes with $\phi_{H,C} = \sqrt{2}$, which is characteristic of patterns behaving as entropy and complexity hubs. We can observe in all cases that $C$ faces the most amplification (or attenuation) of its values (see maximum and minimum values in figure 5), while entropy amplification is always bounded between 0.9 and
Figure 5. Amplifications $A_H \times A_C$ plane. Solid lines represent the model fit. Dashed lines define the regions of interest $R_1$, $R_2$, $R_3$ and $R_4$ allocating nodes of different characteristics. The coefficients of determination are $R^2 = 0.66$ for dengue (a), $R^2 = 0.61$ for influenza (b) and $R^2 = 0.81$ for malaria (c). (Online version in colour.)

1.15. In region $R_2$, we observe how dengue (figure 5a) has three nodes that increase around three times their incoming complexity, while slightly reducing their entropy. However, it is influenza (figure 5b) that has the node with the highest complexity amplification, which increases the incoming complexity up to four times. On the other hand, malaria (figure 5c) is the disease where the amplification of complexity is the lowest, with only one node with a value higher than 2.

In contrast, $R_4$ allocates nodes that decrease the level of complexity of their incoming patterns and, at the same time, increase their entropy. We can observe that in both dengue and influenza, those nodes with the highest entropy amplification $A_H$ depart from the linear behaviour that seems to exist in the interplay between $A_C$ and $A_H$. In fact, when looking at the linear regression, the coefficient of determination for malaria is higher than that for dengue and influenza, but this behaviour can be attributed to the deviations from the linear trend that are reported at both ends of the distributions. In this way, nodes with higher $A_H$ and $A_C$ behave differently from the rest, since in malaria there are no nodes with extreme values, therefore its coefficient of determination is higher.

For the sake of a complete characterization, we now pay attention to the interplay between the entropy/complexity role of the nodes and their topological importance in the structure of the networks. With this aim, we use the eigenvector centrality ($ec$) to account for the node importance [19]. $ec$ assumes that central nodes are those that are, at the same time, (i) connected to many nodes and (ii) connected to well-connected nodes as well. In parallel with $ec$, we obtain the pattern fluctuation ($f$), which quantifies the variability inside each pattern: those patterns whose elements increase and decrease one after the other will have a high $f$, while those patterns that monotonically increase or decrease will have the lowest $f$ (see Methods). For each disease, assuming a pattern dimension of $D = 4$, we obtain 24 different patterns, whose variability $T$ results in six different levels of fluctuation. The higher the variability $T$, the higher the internal disorder of a pattern $\pi$ and the higher its $f$. Electronic supplementary material, table S1 shows how patterns with $D = 4$ are grouped in six different values of $f$. Figure 6 depicts the fluctuation $f$ versus the
topological importance $ec$, where nodes are grouped in terms of its $f$. We can observe that central (important) nodes are those with low levels of fluctuation. In other words, patterns with low variability are the central ones in the network structure. On the contrary, as the internal fluctuation of a pattern increases, its probability of being a hub decreases. As a consequence, peripheral nodes are associated with patterns with the highest fluctuations.

It is worth noting that, for dengue and influenza, nodes with the lowest fluctuations, i.e. those patterns associated with either a monotonic increase or a monotonic decrease in the number of infected individuals, are those with the highest centrality. This fact is probably caused by the existence of abundant periods when the number of infected individuals monotonically increases or decreases, leading to a high number of appearances of patterns $(0, 1, 2, 3)$ or $(3, 2, 1, 0)$, which increases their number of links and, unavoidably, their eigenvector centrality. At the same time, when a node accumulates a significant part of the centrality, it relegates the rest of the nodes of the network to a secondary role. However, malaria behaves in a different way, since we can observe how the heterogeneity of the eigenvector centrality is not as high as in the other two diseases.

Next, we compare the previous results about the interplay between the node role and its centrality with those obtained with synthetic time series. The reason for this is to qualitatively observe whether dynamical signatures of these diseases are similar to those of classical models. With this aim, we simulated six models of different levels of complexity and entropy. Specifically:

— A linear Gaussian process (LGP, $\mathcal{N}(0, 1)$), with the aim of having a signal with the highest entropy, and, as a consequence, with the lowest complexity.

Figure 6. Interplay between the entropy/complexity role of the nodes and their centrality in disease networks. (Online version in colour.)
— A second-order auto-regressive model (AR(2), $x_{t+2} = 0.7x_{t+1} + 0.2x_t + \epsilon_t$), with $\epsilon_t$ being a Gaussian noise.
— A self-exciting threshold AR model (SETAR($k$; $p_1$, $p_2$)), with $k$ as the number of regimes, and $p_1$, $p_2$ the order of the autoregressive parts. SETAR models have been largely used when modelling ecological systems that oscillate between two nonlinear regimes with different delays. This model accounts for higher levels of complexity by decreasing its entropy levels. We modelled it by the SETAR(2;2,2), switching between regimes $(0.62 + 1.25x_{t-1} - 0.43x_{t-2} + 0.0381\epsilon_t)$ if $x_{t-2} \leq 3.25$ and $(2.25 + 1.52x_{t-1} - 1.24x_{t-2} + 0.0626\epsilon_t)$, otherwise.

We also modelled three chaotic systems:
— a logistic map: $x_{t+1} = 4x_t(1 - x_t)$;
— a Rössler system: $\dot{x} = -y - z$, $\dot{y} = x + 0.2y$, $\dot{z} = 0.2 + z(x - 5.7)$;
— a Lorenz system: $\dot{x} = 10(y - x)$, $\dot{y} = x(28 - z) - y$, $\dot{z} = xy - 2.6667z$.

We set the length of the time series generated by each model to $M = 10^4$ after removing the first 1000 samples to avoid possible transients. For the sake of consistency, we applied the log-transformation to our synthetic time series. We reconstructed the respective symbolic networks with $D = 4$ and calculated the centralities and fluctuations of both patterns. Comparing the dynamics with topology diagrams, we detected that the AR(2) and SETAR models behave similarly to the diseases shown in figure 6. We observed how the SETAR model has a considerable gap between the peripheral and central ones (figure 7a), something that was already observed for the influenza networks (figure 6b). At the same time, the AR(2) model shows a smoother decay combined with higher values of centrality for all patterns (figure 7a), as is the case for malaria (figure 6c). In both cases, peripheral nodes are the ones with the highest fluctuations, while hubs correspond to patterns with an internal monotonic increase/decrease.

We now reproduce the $(H, C)$ planes of global measures for all synthetic models and compare them with the experimental results (figure 2). Figure 7b depicts the $(H, C)$ planes accounting...
for the interplay between these two variables and giving interesting information about the global organization of the dynamics of each disease. Here, the properties of the time series are captured with the distribution of all patterns appearing in the time series for $D = 4$ and visualized in the $(H, C)$ phase diagram. All models are enclosed in between the theoretical maximum and minimum complexities described by the black curves. These curves depend solely on the functional form that describes entropy and disequilibrium, as well as the dimension of the space of probabilities associated with the system. Specifically, acquiring the extreme values (maximum or minimum) of complexity is an optimization problem where we compute the set of probability distributions that optimizes the disequilibrium due to a specific value of entropy. This makes the theoretical curves useful when comparing the dynamics of time series of different natures and/or even lengths, in the same $(H, C)$ space but with the same dimensions.

The Rössler (variable $y$), Lorenz (variable $z$) and Logistic chaotic maps appear in the region of positive correlations between $H$ and $C$, which corresponds to high levels of complexity, which are, in turn, correlated with the low levels of entropy. Note that the region before the maximum shown by the theoretical curves can be considered as the division between ordered and disordered states and crucially determines the kind of correlation between entropy and complexity. In the case of autoregressive models, all of them lie in the region of negative correlations. SETAR depicts a high level of complexity with a tendency to disorder, something that is expected for highly nonlinear coupled systems with Gaussian noise. As soon as noise begins to drive the dynamics, and the nonlinearity vanishes, which is the case for AR(2), complexity decreases while entropy increases. In the extreme case of the LGP, its disorder is the highest and its complexity goes to zero.

Finally, it is important to highlight that fluctuations are valid to achieve the types of $H \times C$ diagrams, as well as for identifying the types of nodes. In the electronic supplementary material, we reproduce the results of the relationships between $(C_{in}$ versus $C_{out})$, $(\phi_{H}$ versus $\phi_{C})$ and $(A_{H}$ versus $A_{C})$. Without loss of generality, the figures in the electronic supplementary material are consistent with the behaviour observed in figures 3–5.

4. Discussion

In this work, we investigated the problem of how entropy and complexity flow along the temporal evolution of different epidemic diseases. Time series containing the number of individuals infected with dengue, influenza and malaria are encoded by ordinal patterns, i.e. symbols that are connected to each other sequentially in order to build symbolic graphs. To measure how entropy and complexity flow along ordinal patterns, we endowed symbolic networks with patterns (nodes) that have the ability to transfer their amounts of entropy/complexity to other patterns. Consequently, nodes, or ordinal patterns, might amplify or attenuate the entropy/complexity transmitted to other nodes.

In this way, we propose a family of five parameters to quantify how the entropy/complexity flows among temporal patterns. Two of them quantify how the information arrives (departs) at (from) a pattern in terms of its entropy and complexity. They are normalized, allowing quantitative comparisons between different diseases and synthetic models. Two other features quantify both the level of conductance (flux) and the gain of information of a pattern (amplification). The last one (fluctuation) assesses the inner variability of the dynamics inside each pattern.

Using these metrics, we characterize disease outbreaks, converting epidemiological cohorts into symbolic networks, where temporal fluctuations in signals might communicate among each other. The aim of this approach is to demonstrate whether the exchange of information among each signal’s patterns shows differences among different diseases.

One may consider the incoming/outgoing complexity as information that enters/leaves a given pattern that, in turn, consists of a dynamical state of the disease prevalence given by the combination of $D$ sequential values, accounting for the number of infected individuals. In this context, we have seen in figure 3 how the incoming and outgoing complexities are positively correlated for the three diseases. In this way, the more information a signal’s segment receives
from a previous one, the more information it sends to the following temporal structures. Of the three diseases, influenza is the one with the highest correlation. By contrast, malaria is the disease with the lowest correlation, suggesting the influence of other variables in the transmission of complexity levels.

To the best of our knowledge, this work presents the first evidence of the relationship between the inner dynamics of ordinal patterns, in terms of their fluctuations, and the role they play in the network structure in terms of their centrality variations. Thus, fluctuation \( f \) seems an interesting quantity when measuring the internal dynamics of temporal structures in a signal. By associating node fluctuations with their structural importance, we discover how hub patterns correspond to those nodes with the lowest fluctuation (i.e. hubs are nodes whose related patterns behave more monotonically). Reciprocally, the more peripheral a node is, the more fluctuations the patterns have. Interestingly, this behaviour is reported in all diseases. However, the distribution of centrality in malaria is more homogeneous, with the centrality of both hubs and peripheral nodes being between 0.1 and 0.4 (figure 3c). On the other hand, the distribution of centrality in dengue and, especially, influenza is quite heterogeneous, with few nodes acquiring a centrality above 0.5 (figure 3a,b). This ‘accumulation’ of centrality might be driven by the seasonal nature of influenza, and seasonal-like nature of the dengue data for some of the countries considered. Nodes with low \( f \) appear regularly in the time series. Importantly, when nodes are grouped in terms of their variabilities (see the electronic supplementary material), fluctuation shows comparable results to those that were disclosed by flux, amplitude and nodal complexity. For synthetic time series, we detected qualitative similarities between influenza and nonlinear autoregressive models, while observing that malaria resembles more closely the linear AR processes; these relationships were confirmed by a simple analysis of the entropy–complexity plane.

It is worthwhile highlighting that the family of novel nodal features is valid not only for symbolic networks, but also for any kind of weighted–directed graphs. With directed and weighted networks, e.g. in metabolic, genetic or systems biology scenarios, one can translate them to directed graphs by repeating the methodology proposed in this paper.

Nonetheless, further research is needed to better understand both the structure and dynamics of symbolic networks. On one hand, the construction of symbolic networks from time series should be evaluated when patterns contain delays \( \tau \neq 1 \). Additionally, working with large time series is desirable and would allow larger dimensions \( D \) for the pattern lengths to be obtained. Hence, obtaining large size connected networks would benefit their statistical properties. Apart from that, a detailed analysis of the robustness of this kind of network remains and would help to understand the global properties of these particular graphs.

On the other hand, one of the weaknesses of our study was the poor quality of the available data. Unfortunately, the lack of organization in the acquisition and storage of surveillance data from developing countries makes it difficult to obtain deeper conclusions for healthy public policies. Datasets usually lack complete information. For example, there exist a few weeks with zero reported cases in countries where it is well known that an infected population is continually present. In addition, epidemiological weeks do not reflect the real situation with infected individuals. Because of human migration, infected subjects in one country could have been infected another country (imported cases), which is a problem when performing studies of spatio-temporal differentiation. At the time of retrieving the data, there were no other countries with available datasets for the three diseases we consider, and, surprisingly, influenza is not well documented in most cases. In addition, although dengue and malaria belong to vector-borne diseases, both of them have different strains of virus. Hence, a study with high-quality datasets would find differences among them, but it is mandatory to have large and detailed prevalence records. In general, studies with different datasets would shed light on how different/similar real systems are by means of these entropy/complexity features.

All in all, this work opens the door to new experimental designs to extract information from time series, as well as from directed and weighted networks. However, this methodology should be used as a complementary tool to analyse time series of real datasets. In general, the use of these
metrics and their related methodologies will provide new information to better understand the interplay between temporal structures in natural and artificial collections of finite-size samples.

**Data accessibility.** Disease data are available at: https://github.com/JohannHM/Disease-Outbreaks-Data.

**Authors’ contributions.** J.L.H.-D. and J.H.M. worked on acquisition of the data, J.M.B. and J.H.M. on the conception and experimental design, J.L.H.-D., J.H.M., M.C. and J.M.B. worked on the analysis and interpretation of data, J.H.M. drafted the article. All authors collaborated equally on writing the final manuscript and revising it critically for intellectual content.

**Competing interests.** We declare we have no competing interests.

**Funding.** J.H.M. thanks FAPESP (grant no. 2016/01343-7) for funding a visit to ICTP-SAIFR from March to April 2019, where part of this work was done. J.L.H.-D. was supported by the São Paulo Research Foundation (FAPESP) under grant nos. 2016/01343-7 and 2017/05770-0. J.M.B. is funded by MINECO (project no. FIS2017-84151-P).

**Acknowledgements.** J.H.M. thanks Xavier Bosch-Capblanch, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland, and C. Martínez for valuable conversations.

**References**

1. Granger C, Newbold P. 1986 The theory of forecasting. In Forecasting economic time series (eds C Granger, P Newbold), 2nd edn, pp. 120–150. New York, NY: Academic Press.
2. Liming Y, Guixia EVR, Huajun T. 2013 Time-series modelling and prediction of global monthly absolute temperature for environmental decision making. *Adv. Atmos. Sci.* **30**, 382–396.
3. Noakes DJ, Hipel KW, McLeod A, Jimenez C, Yakowitz S. 1988 Forecasting annual geophysical time series. *Int. J. Forecast.* **4**, 103–115. (doi:10.1016/0169-2070(88)90012-X)
4. Kim S, Kim H. 2016 A new metric of absolute percentage error for intermittent demand forecasts. *Int. J. Forecast.* **32**, 669–679. (doi:10.1016/j.ijforecast.2015.12.003)
5. Weron R. 2014 Electricity price forecasting: a review of the state-of-the-art with a look into the future. *Int. J. Forecast.* **30**, 1030–1081. (doi:10.1016/j.ijforecast.2014.08.008)
6. Noury A, Amini M. 2019 An access and inference control model for time series databases. *Future Gener. Comput. Syst.* **92**, 93–108. (doi:10.1016/j.future.2018.09.057)
7. Rowsseeuw P, Perrotta D, Riani M, Hubert M. 2019 Robust monitoring of time series with application to fraud detection. *Econom. Stat.* **9**, 108–121. (doi:10.1016/j.ecosta.2018.05.001)
8. Si G, Zheng K, Zhou Z, Pan C, Xu X, Qu K, Zhang Y. 2018 Three-dimensional piecewise cloud representation for time series data mining. *Neurocomputing* **316**, 78–94. (doi:10.1016/j.neucom.2018.07.053)
9. Hussain W, Hussain FK, Saberi M, Hussain OK, Chang E. 2018 Comparing time series with machine learning-based prediction approaches for violation management in cloud SLAs. *Future Gener. Comput. Syst.* **89**, 464–477. (doi:10.1016/j.future.2018.06.041)
10. Tak-Chung F. 2011 A review on time series data mining. *Eng. Appl. Artif. Intell.* **24**, 164–181. (doi:10.1016/j.engappai.2010.09.007)
11. Gooijer JGD, Hyndman RJ. 2006 25 years of time series forecasting. *Int. J. Forecast.* **22**, 443–473. (doi:10.1016/j.ijforecast.2006.01.001)
12. Chen G, Abraham B, Bennett GW. 1997 Parametric and non-parametric modelling of time series—an empirical study. *Environmetrics* **8**, 63–74. (doi:10.1002/(SICI)1099-095X(1997018:1%3C63::AID-ENV238%3E3.0.CO;2-B))
13. Kocsis T, Kovács-Székely I. 2017 Comparison of parametric and non-parametric time-series analysis methods on a long-term meteorological data set. *Central Eur. Geol.* **60**, 316–332. (doi:10.1556/24.60.2017.011)
14. He G, Duan Y, Peng R, Jing X, Qian T, Wang L. 2015 Early classification on multivariate time series. *Neurocomputing* **149**, 777–787. (doi:10.1016/j.neucom.2014.07.056)
15. Hoga Y. 2017 Monitoring multivariate time series. *J. Multivariate Anal.* **155**, 105–121. (doi:10.1016/j.jmva.2016.12.003)
16. Aboagye-Sarfo P, Mai Q, Sanfilippo FM, Preen DB, Stewart LM, Fatovich DM. 2015 A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. *J. Biomed. Inform.* **57**, 62–73. (doi:10.1016/j.jbi.2015.06.022)
17. Salles R, Belloze K, Porto F, Gonzalez PH, Ogasawara E. 2019 Nonstationary time series transformation methods: an experimental review. Knowledge-Based Syst. 164, 274–291. (doi:10.1016/j.knosys.2018.10.041)

18. Gubner JA. 2006 Probability and random processes for electrical and computer engineers. Cambridge, UK: Cambridge University Press.

19. Newman M. 2010 Networks: an introduction. Oxford, UK: Oxford University Press.

20. Estrada E, Knight P. 2015 A first course on network theory. Oxford, UK: Oxford University Press.

21. Zanin M, Papo D, Sousa P, Menasalvas E, Nicchi A, Kubik E, Boccaletti S. 2016 Combining complex networks and data mining: why and how. Phys. Rep. 635, 1–44. (doi:10.1016/j.physrep.2016.04.005)

22. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. 2018 Next-generation machine learning for biological networks. Cell 173, 1581–1592. (doi:10.1016/j.cell.2018.05.015)

23. Tanizawa T, Nakamura T, Taya F, Small M. 2018 Constructing directed networks from multivariate time series using linear modelling technique. Physica A 512, 437–455. (doi:10.1016/j.physa.2018.08.137)

24. Wang D, Zhao Y. 2019 Network community detection from the perspective of time series. Physica A 522, 205–214. (doi:10.1016/j.physa.2019.01.028)

25. Zou Y, Donner RV, Marwan N, Donges N. 2019 Complex network approaches to nonlinear time series analysis. Phys. Rep. 787, 1–97. (doi:10.1016/j.physrep.2018.10.005)

26. Lacasa L, Luque B, Ballesteros F, Luque J, Nuño JC. 2008 From time series to complex networks: the visibility graph. Proc. Natl Acad. Sci. USA 105, 4972–4975. (doi:10.1073/pnas.0709247105)

27. Iacovacci J, Lacasa L. 2019 Visibility graphs for image processing. IEEE Trans. Pattern Anal. Mach. Intell. 42, 1–16.

28. Masoller C, Hong Y, Ayad S, Gustave F, Barland S, Pons AJ, Gómez S, Arenas A. 2015 Quantifying sudden changes in dynamical systems using symbolic networks. New J. Phys. 17, 023068. (doi:10.1088/1367-2630/17/2/023068)

29. Lin J, Keogh E, Lonardi S, Chiu B. 2003 A symbolic representation of time series, with implications for streaming algorithms. In Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, New York, NY, 13 June 2003, pp. 2–11. New York, NY: ACM.

30. Collet P, Eckmann J. 1983 Positive Liapunov exponents and absolute continuity for maps of the interval. Ergodic Theory Dyn. Syst. 3, 13–46. (doi:10.1017/S0143385700001802)

31. Kitchen B. 1998 Symbolic dynamics: one-sided, two sided and countable state Markov chains. Berlin, Germany: Springer.

32. Lind D, Marcus B. 1995 An introduction to symbolic dynamics and coding. Cambridge, UK: Cambridge University Press.

33. Hadamard J. 1898 Les surfaces á courbures opposées et leurs lignes géodésiques. J. Math. Pures Appl. 4, 27–73.

34. Morse M, Hedlund G. 1938 Symbolic dynamics. Am. J. Math. 60, 815–866. (doi:10.2307/2371264)

35. Collet P, Eckmann J. 1983 Positive Liapunov exponents and absolute continuity for maps of the interval. Ergodic Theory Dyn. Syst. 3, 13–46. (doi:10.1017/S0143385700001802)

36. Lin J, Keogh E, Lonardi S, Chiu B. 2003 A symbolic representation of time series, with implications for streaming algorithms. In Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, New York, NY, 13 June 2003, pp. 2–11. New York, NY: ACM.
44. Zanin M. 2008 Forbidden patterns in financial time series. Chaos 18, 013119.
45. Cazelles B. 2004 Symbolic dynamics for identifying similarity between rhythms of ecological time series. Ecol. Lett. 7, 755–763. (doi:10.1111/j.1461-0248.2004.00629.x)
46. Rosso OA, Larrondo HA, Martin MT, Plastino A, Fuentes MA. 2007 Distinguishing noise from chaos. Phys. Rev. Lett. 99, 154102. (doi:10.1103/PhysRevLett.99.154102)
47. Health Department, Government of Australia. See https://www.health.gov.au/internet/main/publishing.nsf/Content/Home (accessed 30 September 2017).
48. Ministerio de Salud y Protección Social, Gobierno de Colombia. See https://www.minsalud.gov.co/Paginas/default.aspx (accessed 30 September 2017).
49. Ministry of Health, Labour and Welfare, Japan. See https://www.mhlw.go.jp/english/ (accessed 30 September 2017).
50. Secretaría de Salud, Gobierno de México. See https://www.gob.mx/salud/acciones-y-programas/direccion-general-de-epidemiologia-boletin-epidemiologico (accessed 30 September 2017).
51. Ministry of Health, Singapore. Weekly infectious diseases report. See https://www.moh.gov.sg/resources-statistics/infectious-diseases-bulletin (accessed 30 September 2017).
52. Ministerio del Poder Popular para la Salud, Gobierno Bolivariano de Venezuela. See http://www.vicepresidencia.gob.ve/index.php/tag/ministerio-del-poder-popular-para-la-salud/ (accessed 30 September 2017).
53. Tiana-Alsina J, Torrent M, Rosso O, Masoller C, Garcia-Ojalvo J. 2010 Quantifying the statistical complexity of low-frequency in semiconductor lasers with optical feedback. Phys. Rev. A 82, 013819. (doi:10.1103/PhysRevA.82.013819)
54. Amigo J. 2010 Permutation entropy in dynamical systems: ordinal patterns, permutation entropy and all that. Berlin, Germany: Springer Science & Business.
55. Keller K, Unakafov A, Unakafova V. 2014 Ordinal pattern, entropy and EEG. Entropy 16, 6212–6239.
56. Martínez JH, López ME, Ariza P, Chavez M, Pineda-Pardo JA, López-Sanz D, Gil P, Maestú F, Buldú JM. 2018 Functional brain networks reveal the existence of cognitive reserve and the interplay between network topology and dynamics. Sci. Rep. 8, 10525.
57. Zanin M, Zunino L, Rosso OA, Papo D. 2012 Permutation entropy and its main biomedical and econophysics applications: a review. Entropy 14, 1553–1577. (doi:10.3390/e14081553)
58. Martin MT, Plastino A, Rosso OA. 2006 Generalized statistical complexity measures: geometrical and analytical properties. Physica A 369, 439–462. (doi:10.1016/j.physa.2005.11.053)
59. Rosso OA, De Micco L, Larrondo HA, Martin MT, Plastino A. 2010 Generalized statistical complexity measure. Int. J. Bifur. Chaos 20, 775–785. (doi:10.1142/S021812741002606X)
60. Zunino L, Zanin M, Tabak BM, Pérez DG, Rosso OA. 2010 Complexity-entropy causality plane: a useful approach to quantify the stock market inefficiency. Physica A 389, 1891–1901. (doi:10.1016/j.physa.2010.01.007)
61. Zunino L, Tabak BM, Serinaldi F, Zanin M, Pérez DG, Rosso OA. 2011 Commodity predictability analysis with a permutation information theory approach. Physica A 390, 876–890. (doi:10.1016/j.physa.2010.11.020)
62. Carpi L, Saco P, Rosso O. 2010 Missing ordinal patterns in correlated noises. Physica A 389, 2020–2029. (doi:10.1016/j.physa.2010.01.030)
63. Kullback S, Leibler RA. 1951 On information and sufficiency. Ann. Math. Statist. 22, 79–86. (doi:10.1214/aoms/1177729694)