


RESEARCH

Open Access



# Shannon entropy in time-varying semantic networks of titles of scientific paper

Marcelo do Vale Cunha<sup>1,2\*†</sup> , Carlos Cesar Ribeiro Santos<sup>1†</sup>, Marcelo Albano Moret<sup>1,3†</sup> and Hernane Borges de Barros Pereira<sup>1,3†</sup>

\*Correspondence:

[celaocunha@gmail.com](mailto:celaocunha@gmail.com)

<sup>†</sup>All authors contributed equally to this work.

<sup>1</sup>Programa de Modelagem Computacional, Centro Universitário Senai Cimatec, Av. Orlando Gomes, 1845, 41650-010 Salvador, Brasil

<sup>2</sup>Departamento de Ensino, Instituto Federal da Bahia, R. Gileno de Sá Oliveira, 271 - Recanto dos Pássaros, 47808-006 Barreiras, Brasil  
Full list of author information is available at the end of the article

## Abstract

Recent work has employed information theory in social and complex networks. Studies often discuss entropy in the degree distributions of a network. However, no specific work on entropy exists in clique networks. This work is an extension of a previous study that discussed this topic. We propose a method for calculating the entropy of a clique network and its minimum and maximum values in temporal semantic networks based on titles of scientific papers. In addition, the critical network of moments was extracted. We use the titles of scientific papers published in Nature and Science over ten-year period. The results show the diversity of vocabulary over time, based on the entropy values of vertices and edges. In each critical network, we discover the paths that connect important words and an interesting modular structure.

**Keywords:** Networks of cliques, Shannon entropy, Time-varying graphs, Semantic networks, Network theory

## Introduction

Information theory has evolved in recent decades and has been applied in different fields, such as biology, economics and quantum confined systems (Mousavian et al. 2016; Mishra and Ayyub 2019; Nascimento and Prudente 2018; Brillouin 2013). Recently, some authors have introduced these concepts to measure the information contained in the distribution of degrees and geodesic distances from real networks, or in classical models and semantic networks to classify and differentiate these systems by the heterogeneity of their links (Solé and Valverde 2004; Ji et al. 2008; Viol et al. 2019).

In the study of real networks, modeling the dynamics of the entry and exit of vertices and edges of the networks is necessary. The main models include the modeling of a system by a clique network, e.g., movie actor networks (Barabasi and Albert 1999), co-authoring networks (Newman 2001), concepts networks (Caldeira et al. 2006) and semantic networks (Teixeira et al. 2010; Pereira et al. 2011; Pereira et al. 2016; Grilo et al. 2017). The latter considers the network that is composed of words, concepts or entities with semantic meaning represented by the vertices, with edges that consist of connections between two words that appear in the same unit of meaning, that is, in a sentence (phrase), paragraph or title of the analyzed speech (Pereira et al. 2016; Grilo et al. 2017). Semantic networks

that are modeled by a clique network can provide interesting answers for the study of the organization of human language. Teixeira et al. (2010) proposed the incidence-fidelity (IF) index to obtain a critical configuration of the semantic network of an oral discourse. Cunha et al. (2015) applied this index to networks of scientific paper titles based on publications in high-impact factor journals. The Semantic network of titles (SNT) is formed by the union of titles of publications of a scientific journal, over a given period of time, where the words are vertices of the network and the edges connect words that belong to the same title (Pereira et al. 2011). Within this context, Casteigts et al. (2012) formalized the concept of time-varying graph (TVG).

Despite the growing interest in Shannon entropy, no studies have applied this measure to clique semantic networks. Therefore, this work proposes a method that calculates the vertex and edge entropy of an SNT and their maximum and minimum limits for entropy values according to the initial conditions. The findings can be generalized for any clique network.

This work synthesizes the methodology presented in Cunha et al. (2020) and expands the possibilities of its application and results. The dataset includes the titles from the journals *Nature* and *Science* from 1998 to 2008. The networks are built as a TVG and analyzed using a sliding time window proposed by Cunha et al. (2020) and more explained here. The TVG is then called time-varying semantic network of titles (TVSNT).

In addition to the entropy calculation, the (IF) index (Teixeira et al. 2010) is applied to seek the critical network in prominent time windows, to show the connections between the most important vertices.

The results are explored according to the meanings of these indexes, and comparisons between the two systems are performed from the correlations between the entropy values.

## Background

### Network of cliques

Considering their substantial applicability, clique networks fit the modeling of various social systems. We will provide a brief review of the semantic networks of cliques and the semantic networks based on titles of scientific papers.

**Semantic network of cliques.** According to the definition provided in Grilo et al. (2017) and the premise of (Caldeira et al. 2006), we consider a semantic network of cliques as a system of knowledge representation established by a specific context and imbued with functionality intention, where the vertices are words, concepts or entities with semantic meaning and the smallest unit of meaning is the sentence (e.g., a phrase of a text or discourse, title of a scientific paper, and keywords of a paper) and the edges consist of connections between two words that appear in the sentence.

According to this definition, a word changes its meaning depending on its neighbors in a sentence. Thus, a network is the union of these minor units of meaning, i.e., the cliques union. An increasing number of studies are investigating semantic networks of cliques, i.e., Caldeira et al. (2006) analyzed the structure of meaningful concepts in written discourses; Teixeira et al. (2010) and Lima-Neto et al. (2018) applied semantic clique networks to analyze the relationship between two words that emerge in oral speeches from a critical network, that is, a configuration is obtained using an IF index. In this configuration, the network displays the most information with the least residue (Teixeira et al. 2010); Nascimento et al. (2016) analyzed a semantic network formed by the keywords of a

doctoral thesis in the area of Physics Teaching in Brazil from 1996 to 2006; Andrade et al. (2019) employed the measures of the centralities of degree, proximity and betweenness to understand the coherence and consistency of a proposal for a university program with the subjects' menus and work on the semantic networks of the titles of scientific papers.

**SNT.** An SNT is a semantic network of cliques, where each clique represents one title and its words are clique vertices. Consequently, an edge represents the connection between two words that belong to the same title. Some authors have proposed important study methodologies for SNTs: Pereira et al. (2011, 2016) investigated the topological structure of an SNT of scientific papers as a method to analyze the diffusion efficiency of information, Henrique et al. (2014) employed an SNT to compare the titles of journal papers in mathematics education in English and Portuguese; the work by Cunha et al. (2013) considered a TVSNT and observed an effect on the network memory; Cunha et al. (2015) applied the IF index in SNTs of 15 high-impact factor journals and identified the correspondent critical network for each journal; Pereira et al. (2016) examined the evolution of density during the construction of semantic networks as an indicator of the diversity of scientific journal concepts; and Grilo et al. (2017) proposed a method that analyzes the robustness of an SNT using vertex removal strategies, which enable the identification of a critical removal fraction for which the topological structure of the network is changed.

Note that the authors of (Pereira et al. 2011) were the pioneers in the study of SNTs. The authors proposed rules for manual treatment and a method for data collection, construction and analysis of networks. The work by Fadigas and Pereira (2013) uses the same dataset to apply specific indexes for clique networks, which they proposed, and topologically characterizes the networks using these indexes.

**Indices used in this paper.** For each title network, the properties of the clique networks were utilized (Fadigas and Pereira 2013), as shown in Table 1.

### IF index

Based on the premise of (Caldeira et al. 2006), words that occur together in the same sentence were associatively evoked to construct the idea to be presented. According to (Teixeira et al. 2010), based on this criterion, peers whose association is not significant were included in the network and mask the structure formed by the strongest associations. In this way, filtering is necessary to ensure that only the most relevant associations for the discourse are considered in the construction of the network.

To filter a clique semantic network and obtain the optimal network, Teixeira et al. (2010) created the (IF) index, as shown in Eq. 3. IF index generates a network with a critical configuration that contains the maximum amount of information with the minimum amount of textual residue. This index measures how "strong" and "faithful" the relationship between a pair of words is. For a given pair of words, the index considers the frequency of appearance in the text (incidence  $I$ , Eq. 1) and the frequency of appearance in the context, in which at least one word of the pair is evoked (fidelity  $F$ , Eq. 2). The IF index is the product of these two indices, as shown in Eq. 3.

$$I_{(\alpha,\beta)} = \frac{|C_\alpha \cap C_\beta|}{|\bigcup_1^n C_i|} = \frac{S_{(\alpha,\beta)}}{n_q} \quad (1)$$

**Table 1** Main indices of clique network used in this paper

Index	Description
$n_q$	Number of titles in the initial configuration.
$n$	Number of network vertices in the final configuration.
$m$	Number of edges in the final configuration.
$m_0$	Number of edges in the initial configuration.
$n_0$	Number of vertices in the initial configuration, $n_0 \geq n$ .
$\#(v_i)$	Frequency of vertex $i$ in the initial configuration, i.e., the number of titles that contain vertex $i$ ( $1 \leq \#(v_i) \leq n_q$ ).
$\#(i, j)$	Frequency of edge $(i, j)$ in the initial configuration, i.e., the number of titles that contain the words $i$ and $j$ , $1 \leq \#(i, j) \leq n_q$ , and $i, j = 1, 2, \dots, n$ , with $i \neq j$ and $(i, j) = (j, i)$ .
$q_i$	Title size $i$ . Number of vertices of title $i$ in the initial configuration, ( $1 \leq i \leq n_q$ ).
$q_{min}$	Number of vertices of the smallest title in the initial configuration, ( $1 \leq q_{min} \leq n$ ).
$q_{max}$	Number of vertices of the largest clique in the initial configuration, ( $1 \leq q_{max} \leq n$ ).
$\langle k \rangle$	$\langle k \rangle = \frac{\sum_1^n k_i}{n} = \frac{2m}{n}$ , where $\langle k \rangle$ is the average degree of an undirected network and $k_i$ is the degree of a vertex $i$ , that is the number of edges incident on the vertex $i$ .
$k_i^{hub}$	$k_i^{hub} \geq \langle k \rangle + 2\sigma$ , are the degree values of the hubs, that is, vertices of very high degrees. $\sigma$ is the standard deviation of the degree distribution.

"Initial configuration" is related to the isolated cliques, and "final configuration" is related to the built "network of cliques". The indices are valid for each time window considered

$$F_{(\alpha, \beta)} = \frac{|C_\alpha \cap C_\beta|}{|C_\alpha \cup C_\beta|} = \frac{S_{(\alpha, \beta)}}{S_\alpha + S_\beta - S_{(\alpha, \beta)}} \tag{2}$$

$$IF_{(\alpha, \beta)} = I_{(\alpha, \beta)} \times F_{(\alpha, \beta)} = \frac{(S_{(\alpha, \beta)})^2}{n_q \times (S_\alpha + S_\beta - S_{(\alpha, \beta)})} \tag{3}$$

In the Eqs. 1, 2 and 3,  $\alpha$  and  $\beta$  represent the words in a word pair;  $C_i$  is the set of sentences that contain the word  $i$ ; and  $S_\alpha$ ,  $S_\beta$  and  $S_{(\alpha, \beta)}$  are the number of sentences in which the word  $\alpha$ , word  $\beta$  and word pair  $(\alpha, \beta)$ , respectively, appear.  $n_q$  is the total number of sentences in the text. Thus, once IF index is calculated for all pairs of words, its semantic network becomes weighted at the edges.

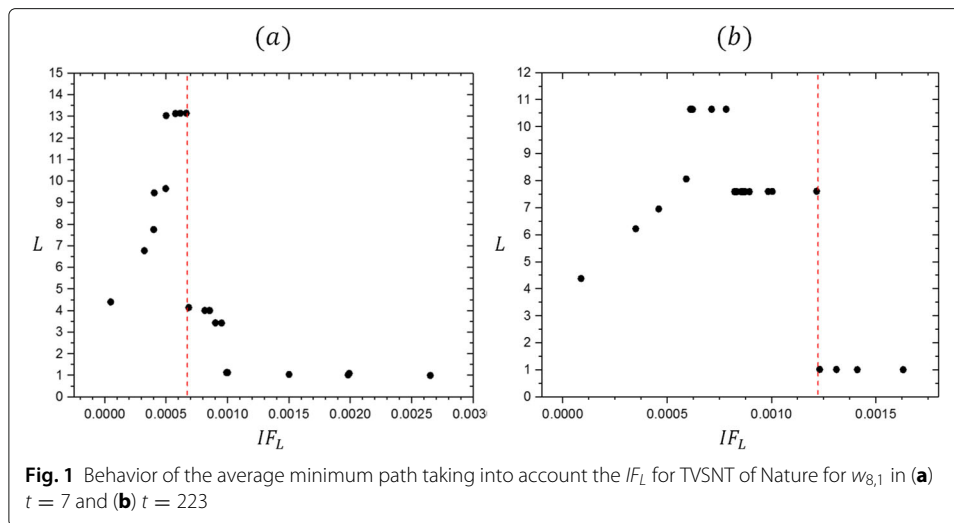
Considering that  $IF_L$  is the minimum allowable value in the network for the IF index, this filtering is performed by removing the edges with  $IF < IF_L$  values; only edges with  $IF > IF_L$  remain in the network.

**Critical Network.** Critical networks were employed to investigate mechanisms inherent to human language in oral speeches (Teixeira et al. 2010; Lima–Neto et al. 2018). A value of  $IF_L = IF_c$  for which the network abruptly changes its connectivity exists. This phenomenon can be verified with the average minimum path in Fig. 1. Figure 2 shows the critical network for the (TVSNT) from scientific papers of Nature,  $w_{8,1}$  in  $t = 8$ .

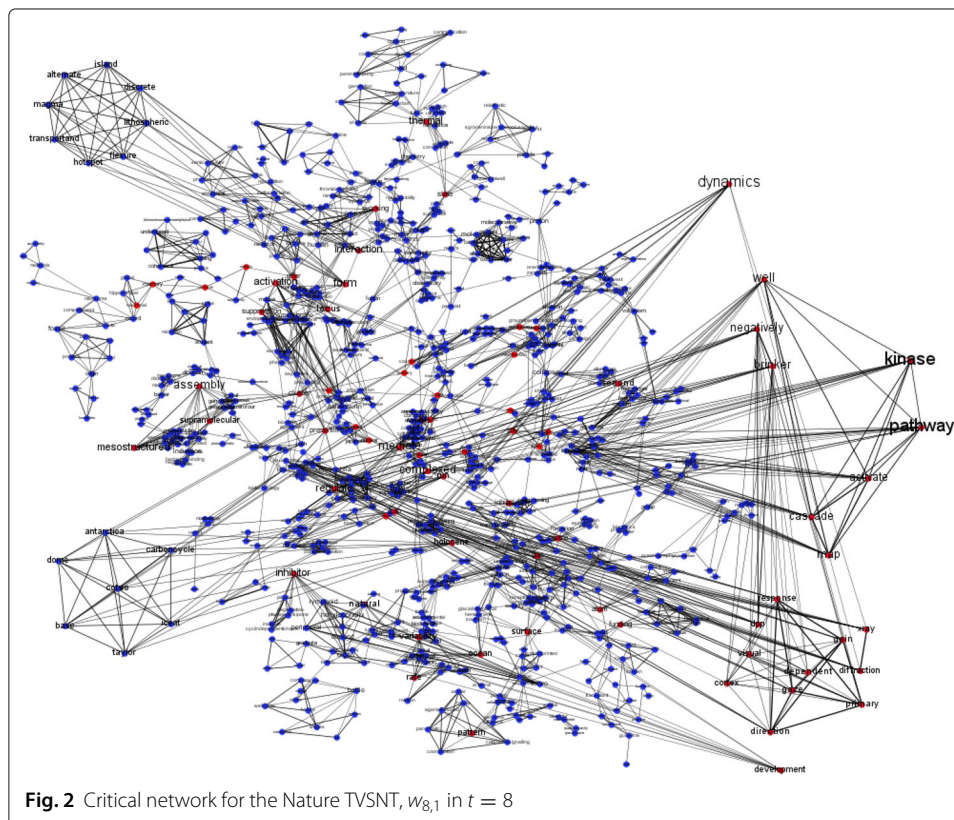
**Temporal networks**

**Brief history.** The use of time is very important in systems analysis in which elements connect. Within the scope of social and complex networks, previous works have been interested in introducing temporal parameters in networks.

Doreian and Stokman (1997) applied models of evolution to study the development of social structures. Barabási et al. (2002) highlighted dynamic and structural mechanisms



in a co-authorship on network and topologically characterized it over time; Li et al. (2007) proposed a model of a scientific collaboration network to verify the scale-free pattern in the weight distributions of the network edges over time. Tang et al. (2010) introduced the concepts of paths and temporal distances and the small word phenomenon in a temporal graph based on the condition of high edge agglomeration and low average temporal distance of nodes in networks of mobile agents and social and biological systems. In 2012, Nicosia et al. (2012) and Casteigts et al. (2012) formalized several concepts and



metrics employed in the study of dynamic networks to create the concept of the TVG, which enables the modeling and analysis of networks that have edges and/or vertices that vary over time. The TVG also enabled the integration of the vast collection of concepts, formalisms and results obtained in previous works (Nicosia et al. 2012).

Amblard et al. (2011) investigated the co-authoring relationships and citations among authors of scientific articles; Silva et al. (2012) analyzed the temporal evolution of brain signals in neuron networks of free-acting rats; Cunha et al. (2013) investigated the memory effect in the time series of a network of titles in the journal Nature; Paranjape et al. (2017) defined temporal network motifs as induced subgraphs on sequences of edges; Holme and Saramäki (2012, 2013) introduced several applications, suggestions for algorithms and specific metrics for networks that vary over time. Holme and Saramäki (2013) discussed the optimal transport structure and relationship between the temporal length and geometric length in a temporal network; Cunha et al. (2020) proposed a method to analyze a TVG from a sliding time-window and build a time series of network indexes, and Sousa et al. (2020) proposed a model named Preferential interaction, which reproduces a weighted-free network of time-varying scale for systems of fixed number of vertices, which can be applied to the investigation of electroencephalogram signal networks in individuals.

**TVG.** Considering the formalization proposed by Nicosia et al. (2012); Casteigts et al. (2012), a TVG is a static graph  $G = (V, E)$  with temporal parameters (functions or sets): presence function ( $\Upsilon$ ), latency function ( $\sigma$ ) and lifetime ( $\Gamma$ ). Thus, a TVG is the fivefold shown in Eq. 4:

$$G = (V, E, \Upsilon, \sigma, \Gamma). \tag{4}$$

In Eq. 4,  $V = \{v_1, v_2, \dots, v_n\}$  is the set of vertices and  $E = \{e_1, e_2, \dots, e_m\}$  is the set of edges of the system, where  $e_k = (i, j)$ , with  $i \neq j$  and  $i, j = (1, 2, \dots, n - 1, n)$ . For these sets,  $n = |V|$  and  $m = |E|$ . The time sets are presented as follows:  $\Gamma \subset \mathbb{N} | \Gamma = \{t_1, t_2, t_3, \dots, t, \dots, t_{(\Psi-1)}, t_\Psi\}$  represents the system lifetime, which is discrete in time. Each element of  $\Gamma$  represents a date or time instant. The interval between the extreme dates is the total time  $T = t_\Psi - t_1 + 1$ .  $\Upsilon = E \times \Gamma \rightarrow \{0, 1\}$  is the presence function that guarantees the existence of a given edge at a given time  $t \in \Gamma$ ; and  $\sigma$  is the latency function, which represents the time required to form an edge.

*Time sliding window function.* The analysis of a TVG can be performed using the sliding window function  $w_{\tau,s}$ , where  $\tau$  is the size of the time window and  $s$  represents the step taken by the window in time (Cunha et al. 2020). Figure 3 shows examples of the use of the function  $w_{\tau,s}$  for a networks analysis.

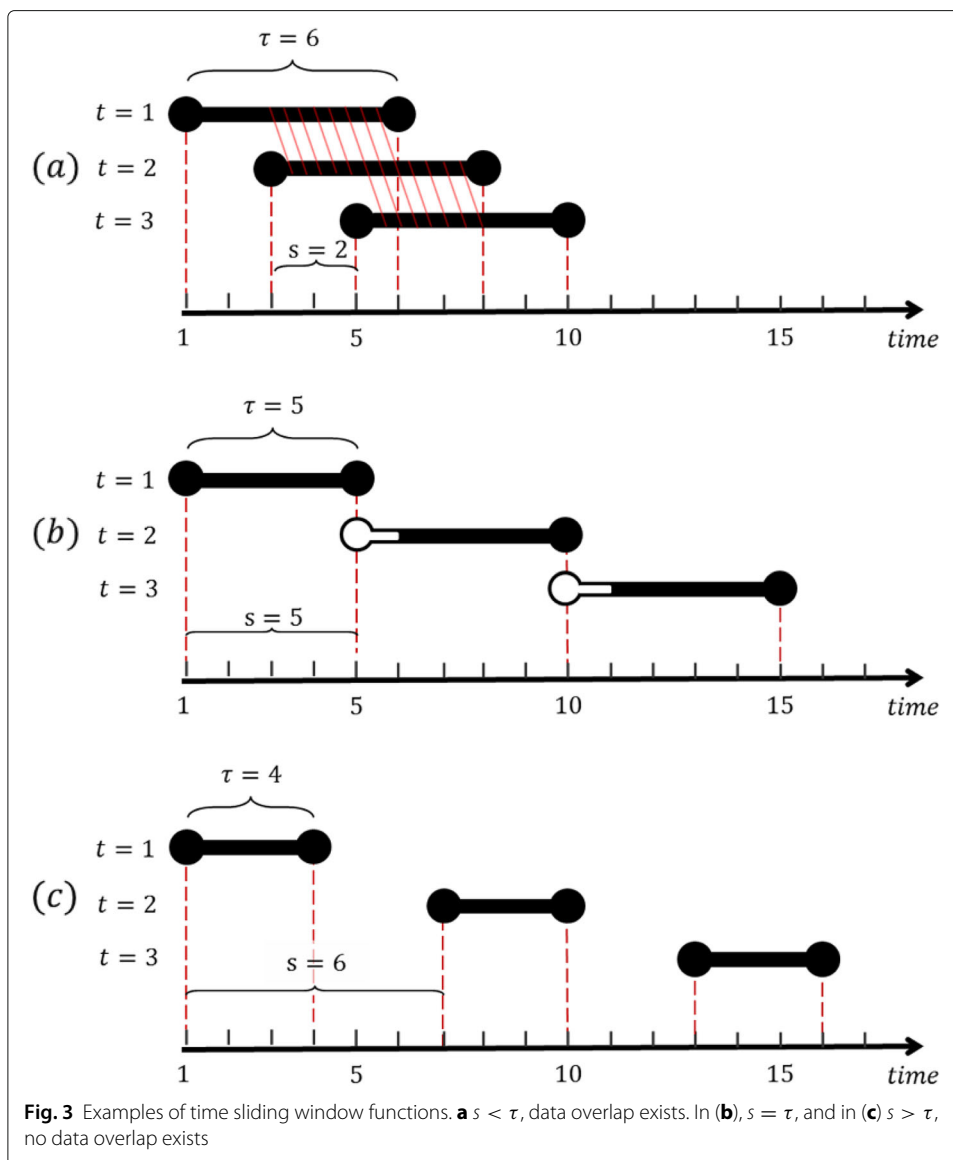
Assuming the values of  $\tau$  and  $s$  are constant and are arbitrated by the researcher, the set of windows fits into the TVG is a function of  $\tau$ ,  $s$  and  $T$ , as shown in Eq. 5. In this equation,  $n_w$  is the number of total windows, i.e., number of networks to be analyzed.

$$w_{\tau,s}(T) = \{\mathcal{G}^1, \mathcal{G}^2, \dots, \mathcal{G}^{n_w-1}, \mathcal{G}^{n_w}\},$$

$$n_w = \left\lfloor \frac{T - (\tau - s)}{s} \right\rfloor. \tag{5}$$

**Information entropy**

The formalism of information as an entropy measure was introduced by Claude Shannon in 1945. According to Shannon theory, the information measure of a variable depends



only on its probability distribution (Shannon 1948). Consequently, the theory can be used in several areas, such as biology, economics, and confined quantum systems (Mousavian et al. 2016; Mishra and Ayyub 2019; Nascimento and Prudente 2018). The theory may compose a methodological link that unites different areas (Zenil et al. 2016), including statistical and thermodynamic physics, in which several recent works have shown some importance for information entropy (Zurek 2018; Gao et al. 2019).

The mathematical concept of information considers that the information contained in a message is associated with the number of possible values or states of this message (Shannon 1948). For example, if the system has only one possible state (e.g., the degree of vertices in a regular network), no information is obtained upon inspection. As the number of possible different states for a system increases, the amount of information in the system increases, that is, the discovery of its real state facilitates further learning.

The entropy is the expected value for the uncertainty of the random variable  $X$  (a system state), which refers to a probability distribution, as shown in Eq. 6.

$$H(X) = -k \sum_i p_i \log p_i \quad (6)$$

In Eq. 6,  $X$  is a random variable,  $p_i$  is the probability of the state  $i$  for this variable (with  $\sum_i p_i = 1$ ), and  $k$  is a constant for which if arbitrated for  $k = \log 2$ , the entropy value is given in bits. The value of  $k$  will be employed. Each calculated entropy value has a maximum value and an associated minimum value. When these limits are known, they help to evaluate how much the real value deviates from these idealized situations.

In a probability distribution for the state of the random variable  $X$ , the minimal entropy situation occurs when the uncertainty is minimal. As an example, when only one possible state for  $X$  exists, we are 100% certain about this state, so  $H(X) = 0$ . The maximum entropy situation occurs when all  $N$  possible states for the variable have an equal probability of occurrence, i.e.,  $p = 1/N$  and  $H(X) = -\sum \frac{1}{n} \log_2(\frac{1}{N}) = \log_2 N$ . Thus, the entropy value for the random variable  $X$  of  $N$  possible states is within these limits, as shown in Eq. 7.

$$0 \leq H(X) \leq \log_2 N \text{ bits} \quad (7)$$

## Method

### Dataset, collection and treatment

The dataset is composed of the titles of articles published in the journals Nature and Science from 1999 to 2008 (Pereira et al. 2011). These journals have high-impact factor values and similar publication frequencies in the collected period<sup>1</sup>.

The words in these titles were treated according to the treatment rules, which were proposed in Pereira et al. (2011) and organized in a way that each week of publications (Journal number) corresponds to a text file, where each line corresponds to a title. The network is then built from these files.

### Building a TVSNT

The SNT is modeled for a TVG, where  $V$  is the set of different words and  $E$  is the set of pairs of words in the same title;  $\Gamma$  is the collected period, which is given in weeks, since a week is the minimum period of publication of the journals. For Nature,  $T = 514$  weeks, and for Science,  $T = 512$  weeks. The presence function  $\Upsilon$  indicates if two words occur in the same title at least once in a given instant. For this work, we will not use the latency function  $\sigma$ , which is a constant.

The sliding time window  $w_{\tau,s}$ , is defined initially as  $\tau = 8$  weeks and  $s = 1$  week, i.e.,  $w_{8,1}$ . The network parameters that are discussed here will be calculated in each window.

### Application information entropy in TVSNT

**Entropy in titles Networks.** According to (Cunha et al. 2020), two random variables can be obtained from the process of titles or cliques network formation: the vertex and the edge. The probabilities of the occurrences of the vertex  $i$  and the edge  $(i, j)$  are calculated for each time window considered, according to Eq. 8 and Eq. 9, respectively. The time

<sup>1</sup>In the period collected, Science has 11798 titles and Nature has 30490 titles, published weekly.



instant  $t$  corresponds to the number of the window.

$$p_i(t) = \left[ \frac{\#(v_i)}{n_0} \right]_t, \text{ with } \sum p_i(t) = 1 \tag{8}$$

$$p_{(i,j)}(t) = \left[ \frac{\#(i,j)}{m_0} \right]_t, \text{ with } \sum p_{(i,j)}(t) = 1 \tag{9}$$

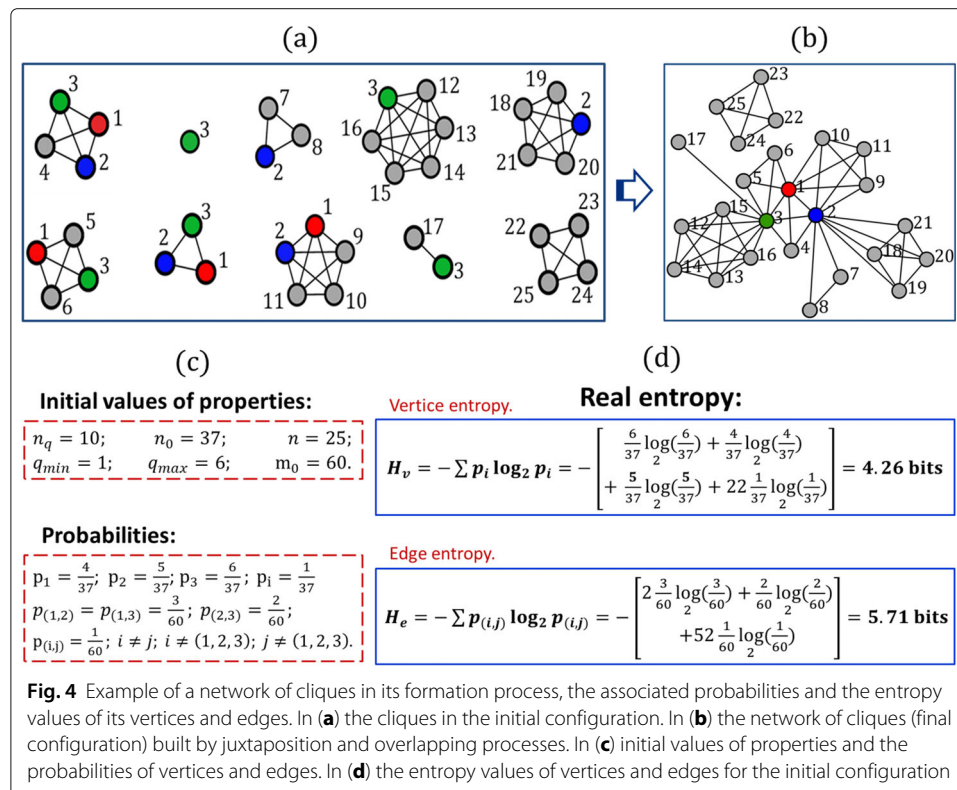
Equations 10 and 11 express the Shannon entropies for these distributions, where  $H_v(t)$  and  $H_e(t)$  represent the entropies of the vertices and edges, respectively, at the given time  $t$ :

$$H_v(t) = - \sum_{i=1}^n p_i(t) \log_2 p_i(t) \tag{10}$$

$$H_e(t) = - \sum_{i \neq j} p_{(i,j)}(t) \log_2 p_{(i,j)}(t) \tag{11}$$

In order to improve the understanding about the calculation of information entropy in TVSNT, we present in Fig. 4 an example of network of cliques and its formation process, the associated probabilities, and the entropy values of vertices and edges. In Fig. 4a, we show the cliques in the initial configuration, and in Fig. 4b, we present the network of cliques built by juxtaposition and overlapping processes (Fadigas and Pereira 2013).

**Limited values for entropy.** The factors that contribute to the increase and reduction of entropy in a system are highlighted here. The minimum entropy value is associated with the variable's maximum certainty. Two factors contribute strongly to this certainty: (i) the minimum of possible states for the variable and (ii) the greater repetition of one or some possible states for the variable.



On other hand, the maximum entropy is associated with the variable's minimum certainty, i.e. with the maximum of possible of states for the variable, where that each state has the lowest possible probability.

The limits shown in Eq. 7 may not apply to the associated entropy from the construction of cliques networks. In this section, the extremes are calculated based on the boundary conditions for the formation of networks.

The following conditions were employed for the investigated journals<sup>2</sup>: the number of cliques in the initial configuration  $n_q$ , size of largest clique  $q_{max}$ , smallest clique size  $q_{min} \neq 0$ , and number of vertices  $n$  and number of vertices in initial configuration  $n_0$ .

To calculate the limits, we will assume the existence of configurations that maximize and minimize the entropy.

**Step 1:** We imagine that the initially empty cliques with  $n_0$  vertices are available to distribute in them, where  $n_0 \geq n$ . Of  $n_0$  vertices,  $n$  is the number of vertices that are necessarily different vertices.

**Step 2:** The  $n$  vertices in the  $n_q$  cliques are distributed without vertex repetition on each clique, where the number of vertices per clique  $q_i$  do not exceed the maximum value  $q_{max}$  and are not less than the minimum value  $q_{min}$ , i.e.,  $q_{min} \leq q_i \leq q_{max}$ .

This moment is referred to as *Configuration 1*. The distribution is performed in a way that there is no repetition of vertices and edges, using Eq. 12. In this configuration, we will have all different vertices and edges with the minimum number of edges. In the final network,  $x$  cliques of size  $q$  and  $y$  cliques of size  $(q + 1)$  exist; thus,

$$\begin{aligned} q &= \left\lfloor \frac{n}{n_q} \right\rfloor \\ y &= n - qn_q \\ x + y &= n_q \\ xq + y(q + 1) &= n \end{aligned} \quad (12)$$

*Configuration 1* generates the highest vertex entropy  $H_{v \max} = \log_2 n$  because it guarantees the disposition of all vertices without repetition and the lowest entropy for the edges of the network once it guarantees the smallest number of edges.

The repetition of a variable also contributes to its reduction in entropy. In clique networks, this phenomenon does not occur for edges because the repetition of an edge implies that the edge exists in more than one clique. Two vertices that compose the edge are forced to be connected to all the other vertices of the clique, which causes a considerable increase in the number of edges, that is, the possibility of an increased number of states, and consequently, an increase in entropy.

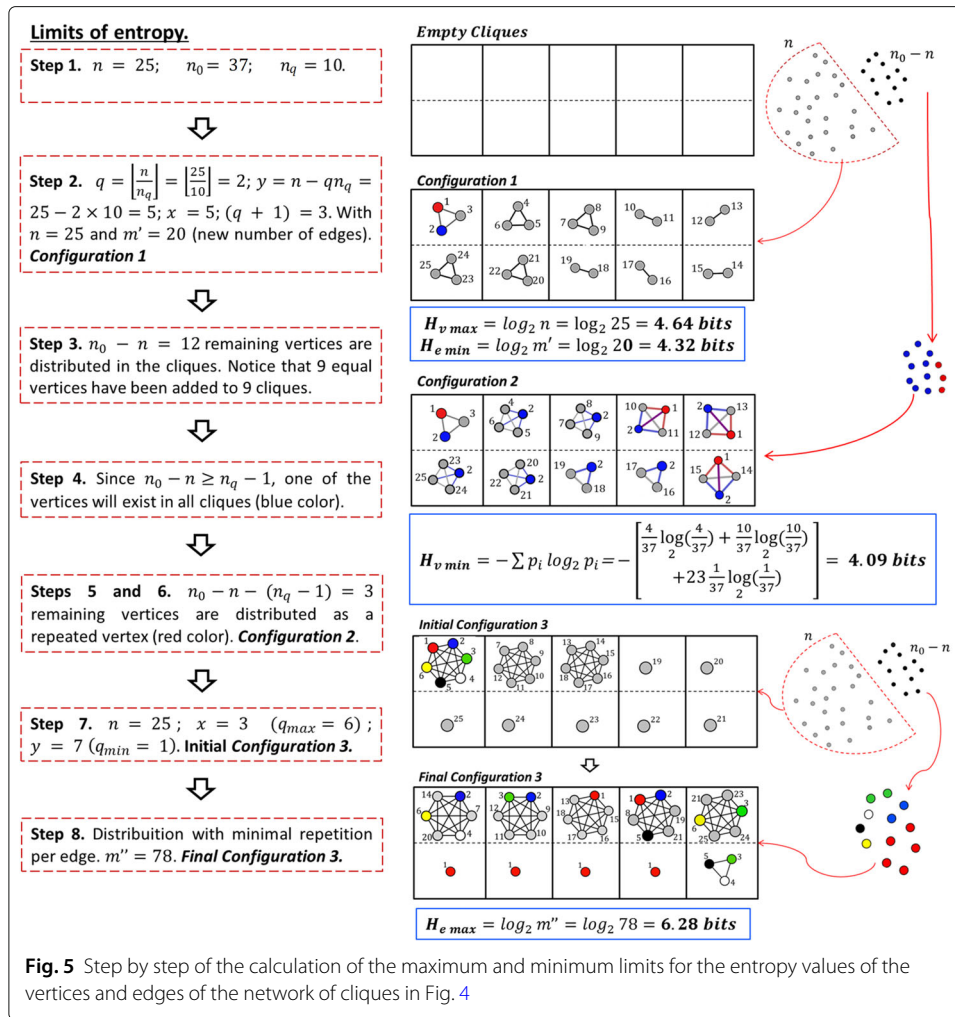
We build *Configuration 2*:

**Step 3:** From *Configuration 1*, the remaining  $n_0 - n$  repeated vertices are added, one by one, with the maximum repetition of vertices for the first vertices added.

**Step 4:** If  $n_0 - n \geq n_q - 1$ , a repeated vertex will exist in all cliques. After the distribution, if  $(n_0 - n) - (n_q - 1) \geq n_q - 1$ , the process continues, with the choice of repeating another vertex in the cliques.

**Step 5:** The process is repeated until the remaining vertices are less than  $n_q - 1$ , and thus, they will be distributed as a single vertex repeated in the number of cliques that can fit.

<sup>2</sup>Depending on the investigated system, the use of all of these conditions or the inclusion or replacement of the existing condition may not be necessary.



**Step 6:** The value  $n_q - 1$  is subtracted from the vertices that have not been added until this subtraction yields a number  $n' \leq n_q - 1$ . Thus, the last vertex is repeatedly added from clique to clique into  $n'$  cliques.

*Configuration 2* increases the probability that some vertices will reduce the entropy to the smallest value possible while respecting the boundary conditions of the problem.

For the maximum edge entropy, the number of edges should be increased as much as possible to avoiding their repetition. For this purpose,

**Step 7:** The appropriate distribution of vertices will be performed according to the initial conditions to obtain a configuration with  $x$  cliques of size  $q_{\max}$  and  $y$  cliques of size  $q_{\min}$ , with the possibility of a clique with size  $q_D$  and  $q_{\min} < q_D < q_{\max}$ , which is referred to as *Initial Configuration 3*.

**Step 8:** The repeated vertices  $n_0 - n$  that remain are separately added to cliques in order to avoid repetition of edges in the cliques (*Final Configuration 3*).

This procedure increases the number of maximum cliques, which causes an increase in the number of distinct edges and, consequently, their entropy.

Using Fig. 4 as a starting point, we summarize in Fig. 5 the process to calculate the maximum and minimum limits for the entropy of vertices and edges.

### Case $n < n_q$

For the TVG of this work, with  $w_{8,1}$ , in every window  $n \geq n_q$ . For larger time windows,  $n < n_q$  may occur. In this case, some adjustments will be required to calculate the limits, for example, in *Configuration 1*,  $q = 0, q+1 = 1, y = n \text{ e } x = n_q - n$ . This case contradicts the boundary condition that  $q = 0 < q_{min}$ . Thus, some  $n - n_0$  will need to be distributed in cliques, in which each clique has the number of vertices  $q = q_{min}$ .

## Results and discussion

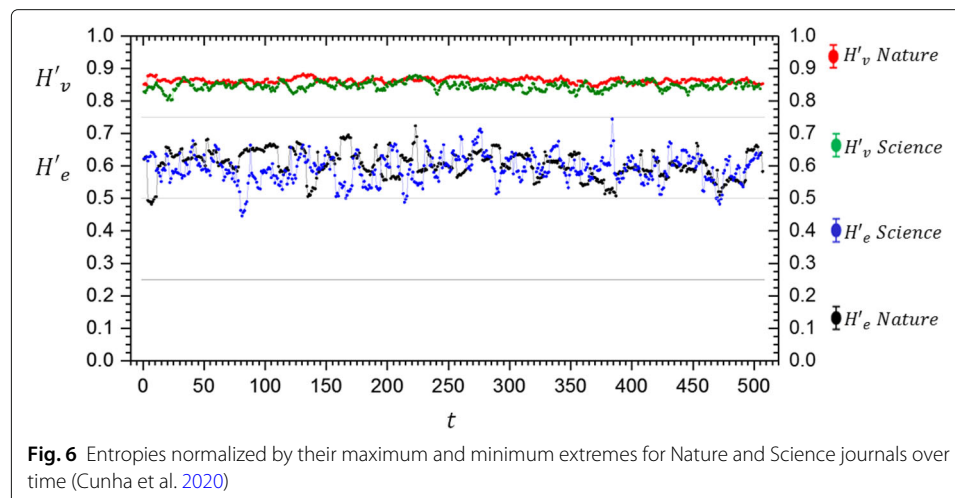
Figure 6 shows the entropy values normalized by their extremes  $H' = \frac{H - H_{min}}{H_{max} - H_{min}}$  of the vertices and edges of the two journals over time. The normalized entropy values eliminate the effect of the size of the networks and allow a better comparison between the diversity of the vocabulary of the titles used to build the semantic networks at different times.

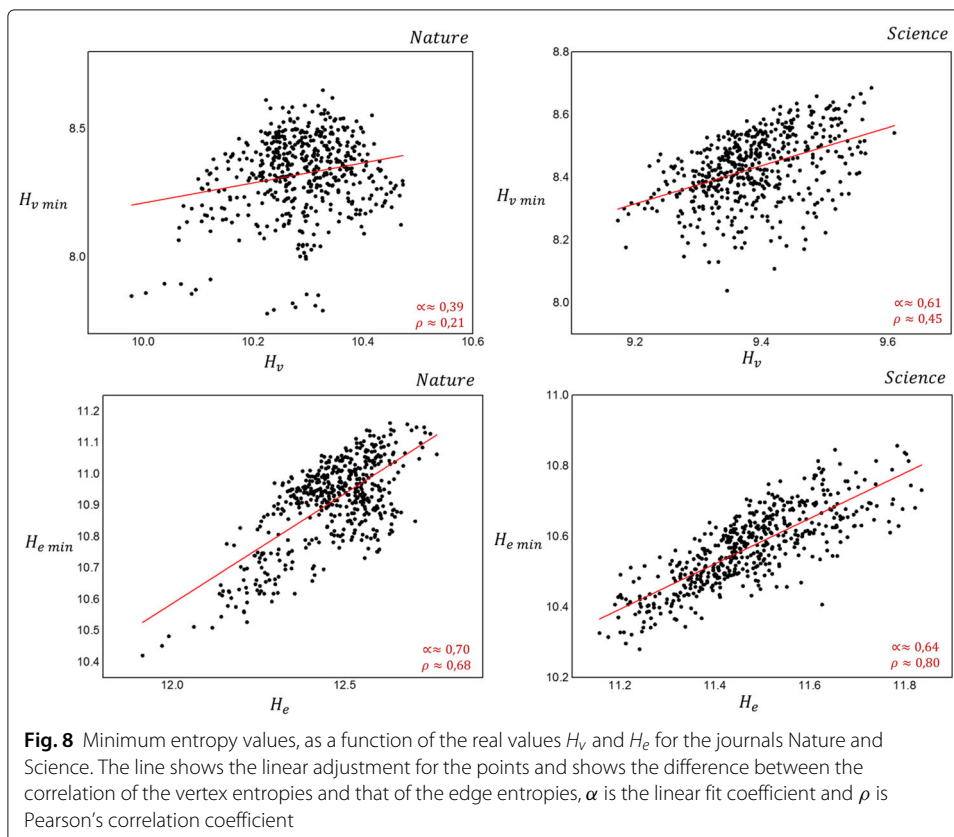
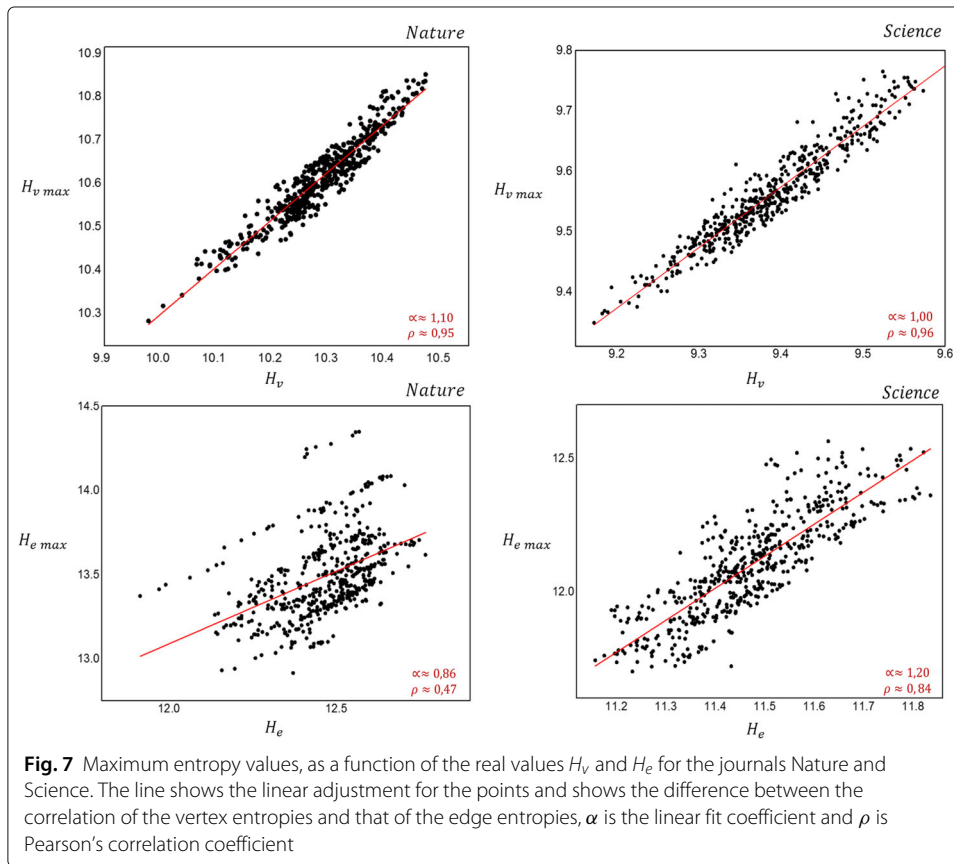
The moments where entropy decreases from its maximum may indicate trends in the journal's vocabulary over time. The vertex entropy values are higher and vary substantially less than the edges entropy values. This finding shows that windows with clique networks have minimal edge overlap.

Moreover, in various intervals,  $H_v$  and  $H_e$  have opposite growth trends. We know that an increasing  $H_e$  implies the generation of new edges, which is possible due to the increment in repeated vertices in the cliques, which causes  $H_v$  to decrease. In some of the study periods, an opposite growth trend was observed between the journals for the edges entropy standard: one journal reached a high entropy value and the other journal had a low entropy value.

Notwithstanding the fact that entropy measures are sensitive to sample size, we use the entire dataset collected in the study period. This approach enables a proper comparison of the two journals, even though they have similar entropy values. Note that the real vertices have entropy  $H_v \cong \log_2 n$  in any time window of the journals. For edge entropy, these values deviate from the corresponding maximum in certain periods. Figures 7 and 8 show how entropies are correlated with their respective maximum and minimum values.

We note a strong correlation between the entropy of vertices and their maximum values, following the entropy of edges with their minimums for both journals. This suggests



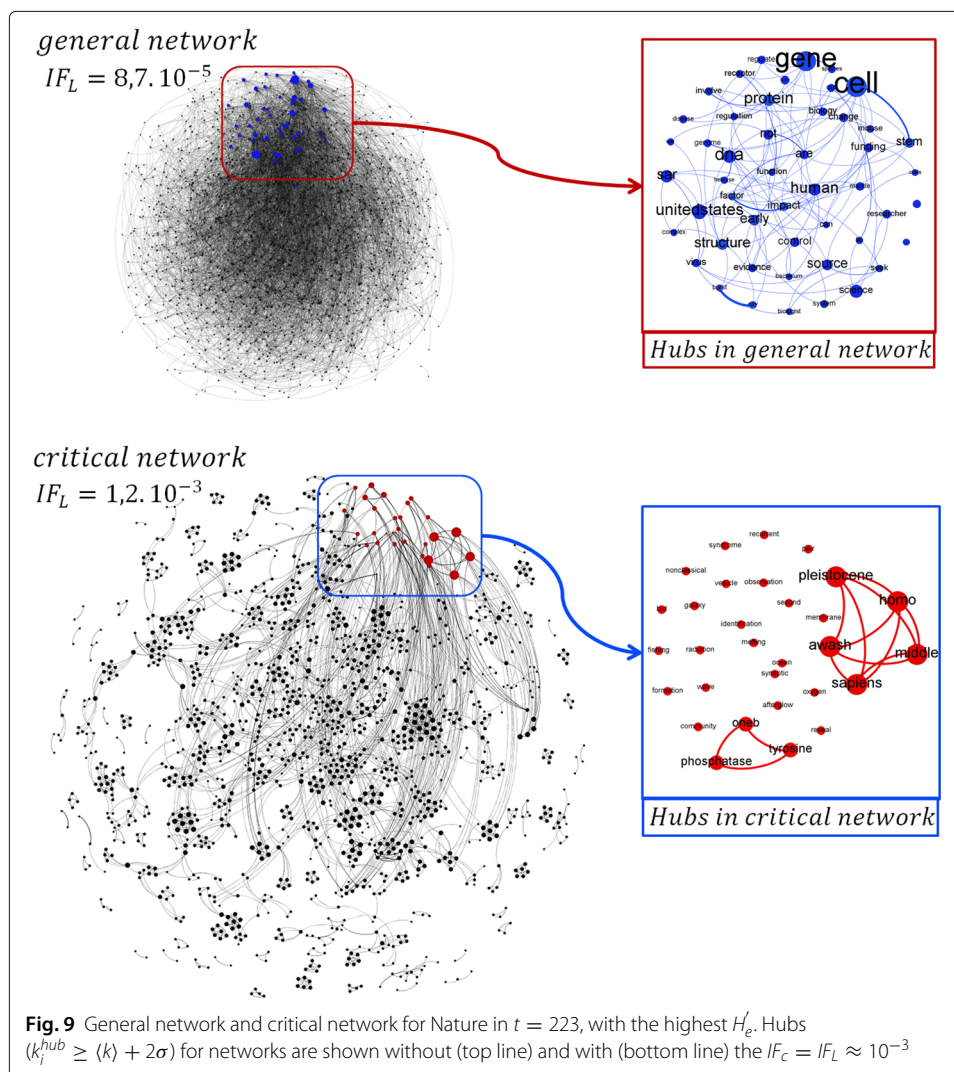


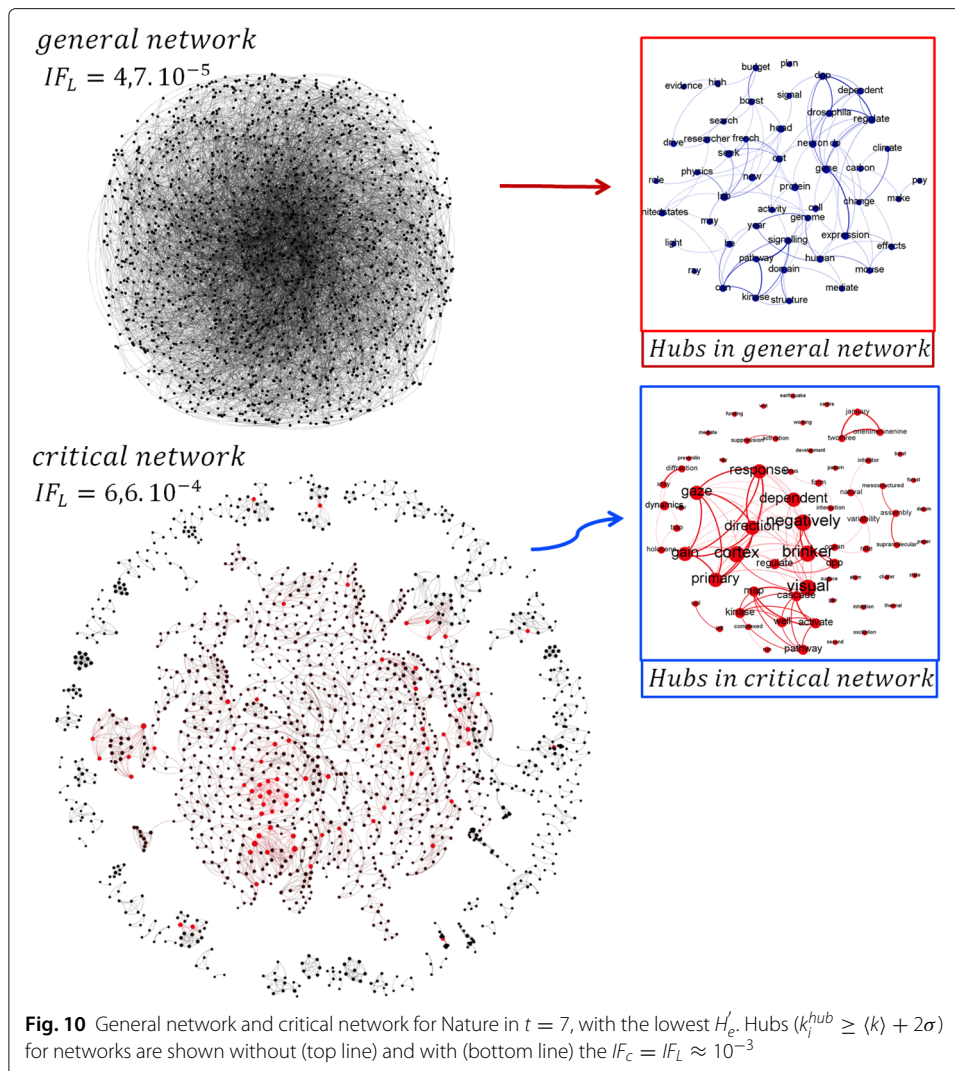
that, over time, the vocabulary of the journals maintained a high diversification for  $w_{8,1}$ , although for  $\tau = T$ , the vocabulary is not that diversified.

The entropy values calculated here do not require the use of a null model (i.e., random network) for comparison. The process of constructing *Configurations 1, 2 and 3* is already randomized. A network of cliques has high clustering, which means that a correspondent random network does not exist since the clustering coefficient tends to zero ( $C \rightarrow 0$ ) in random networks (Watts and Strogatz 1998).

It was found, for each time window, the critical network using the incidence-fidelity index [11]. These networks allowed us to identify the most relevant vertices (i.e. words) considering their connections. Figures 9 and 10 show the critical networks for Nature in  $t = 223$  (the highest  $H'_e$ ) and  $t = 7$  (the lowest  $H'_e$ ) and the vertices considered hubs ( $k_i^{hub} \geq \langle k \rangle + 2\sigma$ ).

As in Teixeira et al. (2010), the TVSNT studied in this work presented a critical network for  $IF_c = IF_L \approx 10^{-3}$ . We highlight, on the one hand, that in the critical network from a network with high entropy, the hubs are poorly connected to each other, indicating greater diversity of the vocabulary. On the other hand, in the critical network from a network





with low entropy, hubs are strongly connected to each other, indicating the robustness and recurrence of vocabulary.

The increase in the entropy measure may be associated with the emergence of new ideas represented by the diversity of the vocabulary and the connections between the words of the titles used to build the semantic network; while the decrease in the entropy measure may be associated with the robustness and consolidation of ideas and interests of authors and editors of a journal in a given time window.

### Conclusions

The results of this study show a strong correlation between the entropy values and their respective maximum values, especially for vertices entropy. It is reasonable to say that it is equivalent to calculate the maximum entropy to estimate the entropy.

Figure 6 shows that journals have a greater diversity of words than word pairs.

With the journal's vocabulary in a window, the number of possible combinations for word pairs is greater than that for repeating them in titles.

When applying the IF index, we noticed that in the critical network, it is possible to identify the main themes, and how they are linked via their vocabulary (specifically, greater diversity of the vocabulary for network with high entropy and robustness and recurrence of the vocabulary for network with low entropy).

The measurement of vocabulary diversity and the diversity of connections between words in a semantic network of scientific article titles allows us to follow (i) the emergence of new ideas over time, represented by the increase in vocabulary diversity of titles or (ii) the robustness and consolidation of ideas and interests of authors and editors of a journal in a given time frame.

The method for constructing clique semantic networks is coherent with previous works with regard to the vocabulary diversity of high-impact scientific journals. The study of vertices and edges entropy in clique networks can be combined with the emergence of communities in these networks and the correlations with other indicators that are specific to this type of network (e.g. reference diameter and fragmentation (Fadigas and Pereira 2013)).

#### Abbreviations

IF: Incidence-fidelity; SNT: Semantic network of titles; TVG: Time-varying graph; TVSNT: Time-varying semantic network of titles

#### Acknowledgements

The authors thank the Federal Institute of Education, Science and Technology of Bahia, Barreiras; the Pro-Rector for Research, Graduate and Innovation (PRPGI - IFBA); and the Senai Cimatec-BA University Center for their financial support.

#### Author's contributions

All authors contributed equally to this work. The author(s) read and approved the final manuscript.

#### Funding

This work also received financial support from National Counsel of Technological and Scientific Development — CNPq, Brazil (grant number 305291/2018-1).

#### Availability of data and materials

The dataset can be easily found and collected on the Journals' websites: <https://www.nature.com/nature/articles> and <https://science.sciencemag.org/content/by/year>.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Programa de Modelagem Computacional, Centro Universitário Senai Cimatec, Av. Orlando Gomes, 1845, 41650-010 Salvador, Brasil. <sup>2</sup>Departamento de Ensino, Instituto Federal da Bahia, R. Gileno de Sá Oliveira, 271 - Recanto dos Pássaros, 47808-006 Barreiras, Brasil. <sup>3</sup>Universidade do Estado da Bahia, Rua Silveira Martins, 2555, Cabula, 41150-000 Salvador, Brasil.

Received: 26 March 2020 Accepted: 26 July 2020

Published online: 26 August 2020

#### References

- Amblard F, Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2011) On the temporal analysis of scientific network evolution. In: CASoN. pp 169–174. <https://doi.org/10.1109/cason.2011.6085938>
- Andrade JC, Barreto RSFD, Cunha MV, Ribeiro NM, Pereira HBB (2019) Interdisciplinaridade e teoria de redes: rede semântica de cliques baseada em ementas e rede de componentes curriculares. *iSys-Revista Brasileira de Sistemas de Informação* 12(3):24–52
- Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
- Barabási A-L, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. *Physica A Stat Mech Appl* 311(3):590–614
- Brillouin L (2013) *Science and Information Theory*. Courier Corporation, North Chelmsford
- Caldeira SMG, Lobão TCP, Andrade RFS, Neme A, Miranda JGV (2006) The network of concepts in written texts. *Eur Phys J B-Condens Matter Complex Syst* 49(4):523–529
- Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2012) Time-varying graphs and dynamic networks. *Int J Parallel Emergent Distrib Syst* 27(5):387–408
- Cunha MV, Miranda JGV, Pereira HBB (2015) Incidência fidelidade aplicada a rede semântica de títulos. In: IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). CSBC, Recife. pp 1–12



- Cunha MV, Rosa MG, Fadigas IS, Miranda JGV, Pereira HBB (2013) Redes de títulos de artigos científicos variáveis no tempo. In: II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). CSBC, Porto Alegre. pp 194–205
- Cunha MV, Santos CCR, Moret MA, Pereira HBB (2020) Shannon entropy in time-varying clique networks. In: H. Cherifi JMEL, Gaito S, Rocha L (eds). International Conference on Complex Networks and Their Applications. Springer, Cham. pp 507–518. Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence
- Doreian P, Stokman F (1997) Evolution of Social Networks. Routledge, London
- Fadigas IS, Pereira HBB (2013) A network approach based on cliques. *Physica A Stat Mech Appl* 392(10):2576–2587
- Gao X, Gallicchio E, Roitberg AE (2019) The generalized boltzmann distribution is the only distribution in which the gibbs-shannon entropy equals the thermodynamic entropy. *J Chem Phys* 151(3):034113
- Grilo M, Fadigas IS, Miranda JGV, Cunha MV, Monteiro RLS, Pereira HBB (2017) Robustness in semantic networks based on cliques. *Physica A Stat Mech Appl* 472:94–102
- Henrique T, Fadigas IS, Rosa MG, Pereira HBB (2014) Mathematics education semantic networks. *Soc Netw Anal Min* 4(1):200
- Holme P, Saramäki J (2012) Temporal networks. *Phys Rep* 519(3):97–125
- Holme P, Saramäki J (2013) Temporal Networks. Springer, Heidelberg
- Ji L, Bing-Hong W, Wen-Xu W, Tao Z (2008) Network entropy based on topology configuration and its computation to random networks. *Chin Phys Lett* 25(11):4177–4180
- Li M, Wu J, Wang D, Zhou T, Di Z, Fan Y (2007) Evolving model of weighted networks inspired by scientific collaboration networks. *Physica A Stat Mech Appl* 375(1):355–364
- Lima-Neto JLA, Cunha MV, Pereira HBB (2018) Redes semânticas de discursos orais de membros de grupos de ajuda mútua: Semantic networks of oral discourses of members of mutual aid groups. *Obra Digital* (14):51–66. <https://doi.org/10.25029/od.2017.177.14>
- Mishra S, Ayyub BM (2019) Shannon entropy for quantifying uncertainty and risk in economic disparity. *Risk Anal* 39(10):2160–2181
- Mousavian Z, Kavousi K, Masoudi-Nejad A (2016) Information theory in systems biology. part i: Gene regulatory and metabolic networks. In: Seminars in Cell & Developmental Biology. Elsevier. pp 3–13. <https://doi.org/10.1016/j.semcd.2015.12.007>
- Nascimento JO, Pereira-Guizzo CS, Moreira DM, Monteiro RLS, Pereira HBB, Moret MA (2016) Redes sociais e complexas: um modelo computacional para a investigação da pós-graduação brasileira em ensino de física. In: VII Encontro Científico de Física Aplicada - Blucher Physics Proceedings. Editora Blucher, São Paulo. pp 110–114
- Nascimento WS, Prudente FV (2018) Shannon entropy: A study of confined hydrogenic-like atoms. *Chem Phys Lett* 691:401–407
- Newman ME (2001) Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys Rev E* 64(1):016132
- Nicosia V, Tang J, Musolesi M, Russo G, Mascolo C, Latora V (2012) Components in time-varying graphs. *Chaos Interdiscip J Nonlinear Sci* 22(2):023101
- Paranjape A, Benson AR, Leskovec J (2017) Motifs in temporal networks. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, Cambridge. pp 601–610
- Pereira HBB, Fadigas IS, Monteiro RLS, Cordeiro AJA, Moret MA (2016) Density: A measure of the diversity of concepts addressed in semantic networks. *Physica A Stat Mech Appl* 441:81–84
- Pereira HBB, Fadigas IS, Senna V, Moret MA (2011) Semantic networks based on titles of scientific papers. *Physica A Stat Mech Appl* 390(6):1192–1197
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27(4):623–656
- Silva BBM, Miranda JGV, Corso G, Copelli M, Vasconcelos N, Ribeiro S, Andrade RFS (2012) Statistical characterization of an ensemble of functional neural networks. *Eur Phys J B* 392:85–358
- Solé RV, Valverde S (2004) Information theory of complex networks: on evolution and architectural constraints. In: E. Ben-Naim ZT, Frauenfelder H (eds). Complex Networks. Springer, Berlin. pp 189–207. Lecture Notes in Physics
- Sousa RA, Lula-Rocha VNA, Toutain T, Rosário RS, Cambui ECB, Miranda JGV (2020) Preferential interaction networks: A dynamic model for brain synchronization networks. *Physica A Stat Mech Appl*. In press
- Tang J, Scellato S, Musolesi M, Mascolo C, Latora V (2010) Small-world behavior in time-varying graphs. *Phys Rev E* 81(5):055101
- Teixeira GM, Aguiar MSF, Carvalho CF, Dantas DR, Cunha MV, Morais JHM, Pereira HBB, Miranda JGV (2010) Complex semantic networks. *Int J Mod Phys C* 21(03):333–347
- Viol A, Palhano-Fontes F, Onias H, de Araujo DB, Hövel P, Viswanathan GM (2019) Characterizing complex networks using entropy-degree diagrams: unveiling changes in functional brain connectivity induced by ayahuasca. *Entropy* 21(2):128
- Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):409–10
- Zenil H, Kiani NA, Tegnér J (2016) Methods of information theory and algorithmic complexity for network biology. In: Seminars in Cell & Developmental Biology, vol. 51. pp 32–43. <https://doi.org/10.1016/j.semcd.2016.01.011>
- Zurek WH (2018) Complexity, Entropy and the Physics of Information. CRC Press, Boca Raton

## Publisher's Note

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