Approximate von Neumann entropy for directed graphs

Cheng Ye,[1,*] Richard C. Wilson,[1,†] César H. Comin,[2,‡] Luciano da F. Costa,[2,§] and Edwin R. Hancock[1,¶]

[1] *Department of Computer Science, University of York, York, YO10 5GH, United Kingdom*

[2] *Institute of Physics at São Carlos, University of São Paulo, PO Box 369, São Carlos, São Paulo, 13560-970, Brazil*

(Received 8 November 2013; revised manuscript received 24 February 2014; published 12 May 2014)

In this paper, we develop an entropy measure for assessing the structural complexity of directed graphs. Although there are many existing alternative measures for quantifying the structural properties of undirected graphs, there are relatively few corresponding measures for directed graphs. To fill this gap in the literature, we explore an alternative technique that is applicable to directed graphs. We commence by using Chung's generalization of the Laplacian of a directed graph to extend the computation of von Neumann entropy from undirected to directed graphs. We provide a simplified form of the entropy which can be expressed in terms of simple node in-degree and out-degree statistics. Moreover, we find approximate forms of the von Neumann entropy that apply to both weakly and strongly directed graphs, and that can be used to characterize network structure. We illustrate the usefulness of these simplified entropy forms defined in this paper on both artificial and real-world data sets, including structures from protein databases and high energy physics theory citation networks.

DOI: 10.1103/PhysRevE.89.052804 PACS number(s): 89.75.Fb, 02.10.Ox, 89.20.Ff, 89.70.Cf

I. INTRODUCTION

Recently, there has been considerable interest in analyzing the properties of complex networks since they play a significant role in modeling large-scale systems in biology, physics, and the social sciences [1–4]. In fact, complex networks provide convenient models for complex systems. Specifically, a complex network is a diagrammatic representation of a complex system. It consists of nodes, which represent the components of the system, and links that connect pairs of nodes, which represent the interconnection between the components. One of the salient characteristics of complex systems is their network structure, i.e., the way in which nodes and links are arranged or organized in a network [5].

To render the analysis of such networks tractable, it is essential to have methods for characterizing their salient properties. One way of viewing complex networks is as graphs whose connectivity properties deviate from those of regular graphs [6]. Whereas regular graphs can be thought of as simple, complex networks are highly nonregular in structure. Structural complexity is therefore perhaps the most important property of a complex network. Computationally efficient measures for quantifying structural complexity are therefore an imperative tool in the analysis of complex networks.

Graph theory offers an attractive route to such methods since it provides effective tools for characterizing network structure together with their intrinsic complexity. This approach has led to the design of several practical methods for characterizing the global and local structure of undirected graphs [7,8]. Unfortunately, there is relatively little literature aimed at studying the structural features of directed graphs. One of the reasons for this is that the graph theory underpinning directed graphs is less developed than that for undirected graphs. In fact, in the real world, those complex networks represented by directed graphs are perhaps even more common than those represented by undirected graphs. For instance, the World Wide Web is a directed network in which nodes represent web pages and links are the hyperlinks between pages. Another common example is furnished by citation networks in which the nodes are scholarly papers while the links are the citations between them.

This paper is motivated by the need to fill this important gap in the literature, and to establish effective methods for measuring the structural properties of directed graphs. In particular, we aim to explore whether the von Neumann entropy previously defined only on undirected graphs [9] can be extended to the domain of directed graphs. To do this, we make use of some recent results from spectral graph theory concerning the construction of the normalized Laplacian matrix for directed graphs [10].

A. Related literature

Quantifying the intrinsic complexity of undirected graphs and networks is a problem of fundamental practical importance, not only in network analysis [11], but also in other areas such as pattern recognition and control theory. Existing approaches are based either on randomness complexity or statistical complexity. The former aims to quantify the degree of randomness or disorganization of a combinatorial structure, while the latter aims to characterize an observed graph structure probabilistically and compute its associated Shannon entropy. Historically, most early work in this area falls into the randomness class, while recent work is statistically based and aims to compute entropic measures of complexity. The reason for this is that randomness complexity does not capture properly the correlations between nodes [12]. Statistical complexity measures regularity and does not necessarily grow monotonically with randomness. A good recent review of the state of the art can be found in the collection of papers edited by Dehmer and Mowshowitz [13].

*cy666@york.ac.uk
†richard.wilson@york.ac.uk
‡appdnails@gmail.com
§ldfcosta@gmail.com
¶edwin.hancock@york.ac.uk

However, while the problem of computing the entropy of undirected graphs is well studied, the literature on directed graphs is rather limited. One recent exception is the work of Riis [14] who has extended the computation of entropy to directed graphs, using the concepts of guessing number and shortest index code. He shows that the entropy is the same as the guessing number and can be bounded by the graph size and shortest index code length. Berwanger *et al.* [15] have proposed a new parameter for the complexity of infinite directed graphs by measuring the extent to which cycles in graphs are intertwined. Recently, Escolano *et al.* [16] have extended the heat diffusion-thermodynamic depth approach for undirected networks to directed networks and thus obtain a means to quantify the complexity of structural patterns encoded by directed graphs.

For undirected graphs, the normalized Laplacian has proved to be a convenient way to represent graph structure that is both linked to the continuous time random walk on a graph and the notion of heat flow on a graph. In fact, thermodynamic depth has proved to provide a powerful means of characterizing a graph in terms of statistical complexity [17]. Recently, the normalized Laplacian spectrum has been shown to provide a complexity level characterization via definition of the von Neumann entropy (or quantum entropy) associated with a density matrix [9,18]. By mapping between discrete Laplacians and quantum states [19], provided that the discrete Laplacian [20] is scaled by the inverse of the volume of the graph, a density matrix is obtained whose entropy can be computed using the spectrum of the discrete Laplacian. For instance, the measure can distinguish between different structures in extremal graph theory. The entropy obtained is maximal for random graphs and is minimal for regular graphs. Han *et al.* [21] have taken this work further and have shown how to approximate the calculation of von Neumann entropy in terms of simple degree statistics rather than the normalized Laplacian eigenvalues. The resulting expression is quadratic in the number of nodes in a graph.

B. Paper outline

The aim in this paper is to explore whether this work on using the von Neumann entropy to characterize the complexity of a graph can be extended from undirected to directed graphs.

One natural way of capturing the structural complexity of directed graphs is to use simple statistics that quantify the balance of in-degree and out-degree at different nodes. A similar but largely heuristic approach has been used to characterize undirected graphs in terms of node degree. In fact, the work of Han *et al.* [21] puts this work on a firmer footing by showing how simple node degree statistics can be used to approximate the von Neumann entropy for undirected graphs. This is a natural step since in information theory, entropy is a measure of unpredictability or information content in a random variable [22]. By extending this definition to graphs, we arrive at a natural way of characterizing their structural complexity. In particular, we can use ideas related to random walks on directed graphs to compute their entropy, and these lead naturally to a characterization in terms of node in-degree and out-degree statistics.

We commence from the work of Passerini and Severini [9], which interprets the normalized Laplacian as a density matrix for an undirected graph, and this in turn allows the graph to be characterized in terms of the von Neumann entropy associated with the density matrix. We extend this work to directed graphs, using Chung's definition of the normalized Laplacian of a directed graph [10]. According to this definition, the directed normalized Laplacian matrix is Hermitian, so the density matrix interpretation of Passerini and Severini still holds in the domain of directed graphs. Furthermore, the von Neumann entropy is essentially the Shannon entropy associated with the normalized Laplacian eigenvalues. Following Han *et al.* [21] we again approximate the Shannon entropy by its quadratic counterpart, with the result that the von Neumann entropy can be simplified in terms of simple in-degree and out-degree statistics. Specifically, the resulting entropy expression depends on the in-degree and out-degree of pairs of nodes connected by links. To further simplify this expression, we consider graphs that are either weakly or strongly directed, i.e., those in which there are large or small proportions of bidirectional links, and develop corresponding approximations of the von Neumann entropy. The approximations accord with our physical intuition concerning in-degree and out-degree on nodes and connecting links.

The outline of this paper is as follows. In Sec. II, we give the detailed development of the simplified forms of von Neumann entropy for directed graphs. In Sec. III, we analyze our theoretical result by undertaking experiments on both artificial and real-world network data. Finally, we conclude the paper with a conclusion of our contribution and suggestions for future work.

II. VON NEUMANN ENTROPY OF DIRECTED GRAPHS

In this section, we propose an entropy measure for characterizing the complexity of directed graphs. Our method is based on extending the definition of von Neumann entropy from undirected to directed graphs. To do this, we commence from Chung's definition of the Laplacian for directed graphs. This leads to an expression for the von Neumann entropy in terms of the in-degree and out-degree statistics of nodes. We then provide approximations of the von Neumann entropy for both strongly directed graphs where there are few bidirectional links and weakly directed graphs where there are few links that are unidirectional.

A. Initial considerations

Suppose $G(V,E)$ is a directed graph with node set V and link set $E \subseteq V \times V$, then the adjacency matrix A is defined as follows:

$$A_{uv} = \begin{cases} 1 & \text{if} \quad (u,v) \in E\ , \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The in-degree and out-degree of node u are

$$d_u^{\text{in}} = \sum_{v \in V} A_{vu}, \quad d_u^{\text{out}} = \sum_{v \in V} A_{uv}. \tag{2}$$

With these ingredients, the transition matrix P for the directed graph G is defined as

$$P_{uv} = \begin{cases} \frac{A_{uv}}{d_u^{\text{out}}} & \text{if} \quad (u,v) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

According to the Perron-Frobenius theorem, for a strongly connected directed graph, the transition matrix P has a unique left eigenvector ϕ with $\phi(u) > 0$, $\forall\, u \in V$, which satisfies $\phi P = \rho\phi$ where ρ denotes the eigenvalue of P. The theorem also implies that if P is aperiodic, the eigenvalues of P have absolute values smaller than the leading eigenvalue $\rho = 1$. Thus, any random walk on a directed graph will converge to a unique stationary distribution if the graph satisfies the properties of strong connection and aperiodicity. We normalize ϕ s.t. $\sum_{i=1}^{|V|} \phi(i) = 1$, this normalized vector corresponds to the unique stationary distribution. Therefore, the probability of a random walker being at node u is the sum of all incoming probabilities of nodes v satisfying $(v,u) \in E$, i.e., $\phi(u) = \sum_{v,(v,u)\in E} \phi(v) P_{vu}$. Based on the properties of the random walk on a directed graph, we assume that the eigenvector component $\phi(u)$ is proportional to the in-degree of the corresponding node d_u^{in}, i.e.,

$$\frac{\phi(u)}{\phi(v)} \approx \frac{d_u^{\text{in}}}{d_v^{\text{in}}}. \tag{4}$$

From this, we derive

$$\frac{\phi(u)}{d_u^{\text{in}}} \approx \frac{\phi(v)}{d_v^{\text{in}}} = \frac{\phi(1) + \phi(2) + \ldots + \phi(|V|)}{d_1^{\text{in}} + d_2^{\text{in}} + \ldots + d_{|V|}^{\text{in}}} = \frac{1}{\text{vol}(G)}, \tag{5}$$

where vol(G) is the volume of graph G, defined as the sum of all node in-degree or out-degree. To illustrate the plausibility of the above assumption, we note that

$$\begin{aligned} \phi(u) &= \sum_{v,(v,u)\in E} \phi(v) P_{vu} = \sum_{v,(v,u)\in E} \frac{d_v^{\text{in}}}{\text{vol}(G)} \frac{A_{vu}}{d_v^{\text{out}}} \\ &= \frac{1}{\text{vol}(G)} \sum_{v,(v,u)\in E} \frac{d_v^{\text{in}}}{d_v^{\text{out}}} = \frac{d_u^{\text{in}}}{\text{vol}(G)} \left\langle \frac{d_v^{\text{in}}}{d_v^{\text{out}}} \right\rangle_{v,(v,u)\in E}. \end{aligned} \tag{6}$$

This implies that the approximation in Eq. (4) holds only when the neighborhood of node u has similar out-degree and in-degree. Although this condition may seem to be a strong requirement, we will undertake experiments in Sec. III to analyze how the local average node degree ratio

$$r_u = \left\langle \frac{d_v^{\text{in}}}{d_v^{\text{out}}} \right\rangle_{v,(v,u)\in E} \tag{7}$$

of u affects the accuracy of our suggested approximate von Neumann entropy (provided later), and the result reveals that this ratio indeed does not cause a significant error.

As stated in Chung [10], if we let $\Phi = \text{diag}[\phi(1), \phi(2), \ldots]$, then the normalized Laplacian matrix of a directed graph can be defined as

$$\tilde{L} = I - \frac{\Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^T \Phi^{1/2}}{2}. \tag{8}$$

Clearly, the normalized matrix is Hermitian, i.e., $\tilde{L} = \tilde{L}^T$ where $\tilde{L}^T$ denotes the conjugated transpose of $\tilde{L}$.

B. Von Neumann entropy of undirected graphs

Passerini and Severini [9] have argued that the combinatorial Laplacian can be interpreted as the density matrix of an undirected graph $G(V,E)$. Therefore, it is possible to define the von Neumann entropy of a graph and calculate it from the eigenvalues of the associated combinatorial Laplacian. In order to gain new insights about the meaning of the von Neumann entropy of a graph, we now show how to obtain a simplified expression for this entropy that can be written in terms of the degrees of the nodes. We commence by summarizing the approximation of the undirected graph von Neumann entropy presented by Han *et al.* [21], and then develop this further to illustrate the limitations of the approximations used.

Although Passerini and Severini have used the traditional Laplacian in their calculations, in order to simplify matters we use the normalized Laplacian $\mathcal{L} = D^{-1/2}(D - A)D^{-1/2}$ (where A is the adjacency matrix and D is the degree matrix with the degrees of the nodes of the undirected graph along the diagonal and zeros elsewhere). In our analysis, the choice of normalization is not an important detail since both Laplacian and normalized Laplacian matrices make valid density matrices. Furthermore, the scaling of the eigenvalues does not affect the functional dependence of the entropy with the degree. In particular, the largest eigenvalue of the Laplacian matrix is bounded by twice the largest node degree in a graph, while the normalized Laplacian matrix has eigenvalues between 0 and 2. With this choice of density matrix, the von Neumann entropy of the undirected graph is the Shannon entropy associated with the normalized Laplacian eigenvalues, i.e.,

$$H_{\text{VN}}^U = -\sum_{i=1}^{|V|} \frac{\tilde{\lambda}_i}{|V|} \ln \frac{\tilde{\lambda}_i}{|V|}, \tag{9}$$

where $\tilde{\lambda}_i, i = 1, \ldots, |V|$, are the eigenvalues of the normalized Laplacian matrix $\mathcal{L}$. Commencing from this definition and making use of the quadratic approximation to the Shannon entropy [i.e., $-x \ln x \approx x(1-x)$, which holds well when x is close to 0 or 1], Han *et al.* [21] approximate the von Neumann entropy by

$$H_Q^U = \sum_{i=1}^{|V|} \frac{\tilde{\lambda}_i}{|V|} \left(1 - \frac{\tilde{\lambda}_i}{|V|}\right). \tag{10}$$

For undirected graphs, this quadratic approximation allows the von Neumann entropy to be expressed in terms of the trace of the normalized Laplacian (which is equal to the sum of the normalized Laplacian eigenvalues) and the trace of the squared normalized Laplacian (which is equal to the sum of the squares of the normalized Laplacian eigenvalues), with the result that

$$H_{\text{VN}}^U = \frac{\text{Tr}[\mathcal{L}]}{|V|} - \frac{\text{Tr}[\mathcal{L}^2]}{|V|^2}. \tag{11}$$

For undirected graphs, the two traces appearing in the above expression are given in terms of statistics for the degrees of nodes in the graph [21], with the result that

$$H_{\rm VN}^U = 1 - \frac{1}{|V|} - \frac{1}{|V|^2} \sum_{(u,v)\in E} \frac{1}{d_u d_v}. \qquad (12)$$

This formula contains two measures of graph structure: the first one is the number of nodes of graph, while the second one is based on degree statistics for pairs of nodes connected by links. Moreover, the computational complexity of this expression is quadratic in graph size, which is much simpler than that of the original entropy.

The accuracy of the above expression depends on the veracity of the quadratic approximation to the Shannon entropy $x \ln x \approx -x(1-x)$. This approximation is known to hold well when either $x \to 0$ or $x \to 1$, which guarantees the accuracy of the quadratic entropy since $\frac{\tilde{\lambda}_i}{|V|} \to 0$ when the graph size is very large.

A more precise expression for the von Neumann entropy can be obtained by making a second-order Taylor series approximation for the Shannon entropy with expansion point x_0 at the mean value of $\frac{\tilde{\lambda}}{|V|}$, i.e.,

$$x_0 = \frac{\sum_{i=1}^{|V|} \frac{\tilde{\lambda}_i}{|V|}}{|V|} = \frac{\mathrm{Tr}[\mathcal{L}]}{|V|^2}.$$

The second-order Taylor expansion for $x \ln x$ about the expansion point x_0 is

$$x \ln x \approx -x\left(-\ln x_0 - \frac{x}{2x_0}\right) - \frac{x_0}{2}.$$

Substituting this series approximation for the Shannon entropy with expansion point

$$x_0 = \frac{\mathrm{Tr}[\mathcal{L}]}{|V|^2} = \frac{1}{|V|}$$

into the expression for the von Neumann entropy [Eq. (8)], we obtain

$$H_T^U = \ln|V| - \frac{1}{2|V|} \sum_{(u,v)\in E} \frac{1}{d_u d_v}.$$

As a result, the Taylor series approximation to the von Neumann entropy at the expansion point $x_0 = \frac{1}{|V|}$ and the quadratic approximation are related by

$$H_T^U = \frac{|V|}{2} H_Q^U + \ln|V| + \frac{1-|V|}{2}.$$

In other words, the two entropies are related by an offset and a scale, which are related to the number of nodes in the graph. Since we are concerned in applying the von Neumann entropy for characterizing the structure of graphs, the differences caused by the influence of graph size do not matter in our analysis. Therefore, both expressions can be used. Throughout the paper, we use the simpler expression given by H_Q^U.

C. Von Neumann entropy of directed graphs

The true von Neumann entropy for a directed graph can be computed using the Shannon entropy associated with the eigenvalues of its normalized Laplacian matrix. Unfortunately, for large graphs this is not a viable proposition since the time required to solve the eigensystem is cubic in the number of nodes. To overcome this problem, we aim to extend the analysis of Han *et al.* [21] from undirected to directed graphs. To do this, we again make use of the quadratic approximation to the Shannon entropy in order to obtain a simplified expression for the von Neumann entropy of a directed graph, which can be computed in a time that is quadratic in the number of nodes. Our starting point is the quadratic approximation to the von Neumann entropy in terms of the traces of normalized Laplacian and the squared normalized Laplacian, i.e.,

$$H_{\rm TVN}^D = \frac{\mathrm{Tr}[\tilde{L}]}{|V|} - \frac{\mathrm{Tr}[\tilde{L}^2]}{|V|^2}. \qquad (13)$$

To simplify this expression a step further, we repeat the computation of traces for the case of a directed graph. This is not a straightforward task, and requires that we distinguish between the in-degree and out-degree of nodes. We first consider Chung's expression for the normalized Laplacian of directed graphs and write

$$\begin{aligned}\mathrm{Tr}[\tilde{L}] &= \mathrm{Tr}\left[I - \frac{\Phi^{1/2}P\Phi^{-1/2} + \Phi^{-1/2}P^T\Phi^{1/2}}{2}\right] \\ &= \mathrm{Tr}[I] - \frac{1}{2}\mathrm{Tr}[\Phi^{1/2}P\Phi^{-1/2}] \\ &\quad - \frac{1}{2}\mathrm{Tr}[\Phi^{-1/2}P^T\Phi^{1/2}]. \end{aligned} \qquad (14)$$

Since the trace is invariant under cyclic permutations, i.e., $\mathrm{Tr}[ABC] = \mathrm{Tr}[BCA] = \mathrm{Tr}[CAB]$, we have

$$\begin{aligned}\mathrm{Tr}[\tilde{L}] &= \mathrm{Tr}[I] - \tfrac{1}{2}\mathrm{Tr}[P\Phi^{-1/2}\Phi^{1/2}] - \tfrac{1}{2}\mathrm{Tr}[P^T\Phi^{1/2}\Phi^{-1/2}] \\ &= \mathrm{Tr}[I] - \tfrac{1}{2}\mathrm{Tr}[P] - \tfrac{1}{2}\mathrm{Tr}[P^T]. \end{aligned} \qquad (15)$$

The diagonal elements of the transition matrix P are all zeros, hence we obtain

$$\mathrm{Tr}[\tilde{L}] = \mathrm{Tr}[I] = |V|, \qquad (16)$$

which is exactly the same as in the case of undirected graphs.

Next, we turn our attention to $\mathrm{Tr}[\tilde{L}^2]$:

$$\begin{aligned}\mathrm{Tr}[\tilde{L}^2] &= \mathrm{Tr}\big[I^2 - (\Phi^{1/2}P\Phi^{-1/2} + \Phi^{-1/2}P^T\Phi^{1/2}) \\ &\quad + \tfrac{1}{4}(\Phi^{1/2}P\Phi^{-1/2}\Phi^{1/2}P\Phi^{-1/2} \\ &\quad + \Phi^{1/2}P\Phi^{-1/2}\Phi^{-1/2}P^T\Phi^{1/2} \\ &\quad + \Phi^{-1/2}P^T\Phi^{1/2}\Phi^{1/2}P\Phi^{-1/2} \\ &\quad + \Phi^{-1/2}P^T\Phi^{1/2}\Phi^{-1/2}P^T\Phi^{1/2})\big] \\ &= \mathrm{Tr}[I^2] - \mathrm{Tr}[P] - \mathrm{Tr}[P^T] + \tfrac{1}{4}\big(\mathrm{Tr}\,[P^2] \\ &\quad + \mathrm{Tr}[P\Phi^{-1}P^T\Phi] + \mathrm{Tr}[P^T\Phi P\Phi^{-1}] + \mathrm{Tr}\big[P^{T^2}\big]\big) \\ &= |V| + \tfrac{1}{2}(\mathrm{Tr}[P^2] + \mathrm{Tr}[P\Phi^{-1}P^T\Phi]), \end{aligned} \qquad (17)$$

which is different to the result obtained in the case of undirected graphs.

To continue the development, we first partition the link set E into two disjoint subsets E_1 and E_2, where $E_1 = \{(u,v)|(u,v) \in E \text{ and } (v,u) \notin E\}$, $E_2 = \{(u,v)|(u,v) \in E \text{ and } (v,u) \in E\}$ that satisfy the conditions $E_1 \bigcup E_2 = E$, $E_1 \bigcap E_2 = \emptyset$. Then, according to the definition of the transition matrix, we find

$$\mathrm{Tr}[P^2] = \sum_{u\in V}\sum_{v\in V} P_{uv}P_{vu} = \sum_{(u,v)\in E_2} \frac{1}{d_u^{\mathrm{out}} d_v^{\mathrm{out}}}. \tag{18}$$

Using the fact that $\Phi = \mathrm{diag}[\phi(1),(2),\ldots]$, we have

$$\mathrm{Tr}[P\Phi^{-1}P^T\Phi] = \sum_{u\in V}\sum_{v\in V} P_{uv}^2 \frac{\phi(u)}{\phi(v)} = \sum_{(u,v)\in E} \frac{\phi(u)}{\phi(v) d_u^{\mathrm{out}2}}. \tag{19}$$

Using Eq. (4), we can approximate the von Neumann entropy of a directed graph in terms of the in-degree and out-degree of the nodes as follows:

$$H_{\mathrm{VN}}^D = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2}\left\{\sum_{(u,v)\in E}\left(\frac{1}{d_u^{\mathrm{out}}d_v^{\mathrm{out}}} + \frac{d_u^{\mathrm{in}}}{d_v^{\mathrm{in}}d_u^{\mathrm{out}2}}\right) - \sum_{(u,v)\in E_1}\frac{1}{d_u^{\mathrm{out}}d_v^{\mathrm{out}}}\right\} \tag{20}$$

or, equivalently,

$$H_{\mathrm{VN}}^D = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2}\left\{\sum_{(u,v)\in E}\frac{d_u^{\mathrm{in}}}{d_v^{\mathrm{in}}d_u^{\mathrm{out}2}} + \sum_{(u,v)\in E_2}\frac{1}{d_u^{\mathrm{out}}d_v^{\mathrm{out}}}\right\}. \tag{21}$$

To take our analysis one step further, it is interesting to explore how the entropy is bounded for graphs of a particular size, and in particular which topologies give the maximum and minimum entropies. From Eq. (21) it is clear that when the terms in the curly brackets reach their largest value, the von Neumann entropy takes on its minimum value. This occurs when the structure is a circle graph, in which each node has only one outgoing link and one incoming link. On the other hand, when the terms in the curly braces take on their smallest value, the entropy is maximum. This occurs when there are no bidirectional links in the graph. Nodes that have outgoing links have no incoming links. A typical example of this type of structure is a star graph.

The maximum and minimum von Neumann entropies corresponding to these cases are as follows. For a circle directed graph $G(V,E)$, all nodes have the same out-degree and in-degree equal to 1, then

$$H_{\mathrm{VN}}^D = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2}|V| = 1 - \frac{1}{|V|} - \frac{1}{2|V|}.$$

Turning attention to the case of a star graph, the center node has out-degree (in-degree) $|V|-1$, and the remaining nodes have in-degree (out-degree) 1. In this case,

$$H_{\mathrm{VN}}^D = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2}\times 0 = 1 - \frac{1}{|V|}.$$

As a result, the approximate von Neumann entropy suggested in Eq. (21) gives the minimum value for the ring graph, which is the simplest regular graph. It takes on its maximum value for star graphs. This latter structure can be viewed as the most complex since it has the greatest difference in node out-degree and in-degree.

To continue the development a step further, we can simplify the approximate von Neumann entropy expression according to the relative sizes of the sets E_1 and E_2, to provide approximations to the von Neumann entropy which are specific to weakly and strongly directed graphs.

For weakly directed graphs, $|E_1| \ll |E_2|$, i.e., few of the links are not bidirectional, and we can ignore the summation over E_1 in Eq. (20). Rewriting the remaining terms in curly braces in terms of a common denominator and then dividing numerator and denominator by $d_u^{\mathrm{out}}d_v^{\mathrm{out}}$ we obtain

$$H_{\mathrm{VN}}^{\mathrm{WD}} = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2}\sum_{(u,v)\in E}\left\{\frac{\frac{d_u^{\mathrm{in}}}{d_u^{\mathrm{out}}} + \frac{d_v^{\mathrm{in}}}{d_v^{\mathrm{out}}}}{d_u^{\mathrm{out}}d_v^{\mathrm{in}}}\right\}. \tag{22}$$

The term $1 - \frac{1}{|V|}$ tends to unity as the graph size becomes large. In the summation, the numerator is given in terms of the sum of the ratios of in-degree and out-degree of the nodes. Since the directed links can not start at a sink (a node of zero out-degree), the ratios do not become infinite.

On the other hand, for strongly directed graphs, there are few bidirectional links, i.e., $|E_2| \ll |E_1|$, and we can ignore the summation over E_2 in Eq. (21), giving the approximate entropy for strongly directed graphs

$$H_{\mathrm{VN}}^{\mathrm{SD}} = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2}\sum_{(u,v)\in E}\left\{\frac{d_u^{\mathrm{in}}}{d_v^{\mathrm{in}}d_u^{\mathrm{out}2}}\right\}. \tag{23}$$

Both the weakly and strongly directed forms of the von Neumann entropy ($H_{\mathrm{VN}}^{\mathrm{WD}}$ and $H_{\mathrm{VN}}^{\mathrm{SD}}$) contain two terms. The first is the graph size, while the second one depends on the in-degree and out-degree statistics of each pair of nodes connected by a link. Moreover, the computational complexity of these expressions is quadratic in the graph size.

There are a number of points to note concerning the development above. First, the normalized Laplacian matrix of directed graphs denoted by $\tilde{L}$ in Eq. (8) satisfies the conditions of Passerini and Severini [9] for the density matrix. Moreover, we have shown that $\tilde{L}$ is Hermitian, so its eigenvalues are all real. Hence, theoretically, the density matrix interpretation of Passerini and Severini [9] can be extended to directed graphs. Second, when the out-degree and in-degree are the same at all nodes, the von Neumann entropy for directed and undirected graphs is identical.

To conclude this section, it is worth discussing the role of sinks in our analysis. A sink is a node with several incident links, but no outgoing links. Hence, they are characterized by zero out-degree. One obvious problem with our formulation is that our expression for the von Neumann entropy of a weakly directed graph will become singular when node v is a sink, i.e., $d_v^{\mathrm{out}} = 0$. However, in the case of weakly directed graphs, the likelihood of sink nodes is small since the probability of bidirectional links is large. We can reach the same conclusion by recalling that the graph represents a Markov chain with equal transition probabilities on the nodes. If the chain is irreducible and aperiodic, then convergence to a stationary distribution is guaranteed. Otherwise, the final distribution may not be stable or may depend on the initial conditions. In particular, if the irreducibility condition is not true, then the Perron-Frobenius theorem does not hold and we can not construct the Laplacian in that case, or at least it is not clear if the theorems in Chung's paper hold. So, if we demand that the Markov chain is irreducible, this means the graph is strongly connected and so there are no sinks in the graph.

III. EXPERIMENTS AND EVALUATIONS

We have derived an expression for the von Neumann entropy of a directed graph, and have provided approximations that apply to both weakly and strongly directed graphs. In this section, we aim to explore whether these entropy measures can be used to determine changes in the structure of directed graphs. We confine our attention to two principal tasks. The first one is to explore whether the entropy measures can be used to distinguish different types of directed graphs. The second is to use the entropy measures to detect abrupt changes in the structure of networks that evolve with time. For most experiments, we normalize the entropy measures studied [including the approximate von Neumann entropy in Eq. (21) together with its approximations for both weakly and strongly directed graphs given in Eqs. (22) and (23), respectively]. The normalization is with respect to the graph size, and this removes some of the size dependence. Specifically, we compute the quantity

$$J_{\mathrm{VN}}^{D} = |V|\left|H_{\mathrm{VN}}^{D} - \left(1 - \frac{1}{|V|}\right)\right| = \frac{1}{2|V|}\left\{\sum_{(u,v)\in E}\frac{d_u^{\mathrm{in}}}{d_v^{\mathrm{in}} d_u^{\mathrm{out}^2}} + \sum_{(u,v)\in E_2}\frac{1}{d_u^{\mathrm{out}} d_v^{\mathrm{out}}}\right\} \tag{24}$$

as a normalized quantity which captures variations in the in-degree and out-degree statistics in the same manner as the approximate von Neumann entropy H_{VN}^{D}. Similarly, the corresponding normalized quantities for the weakly and strongly directed approximations $H_{\mathrm{VN}}^{\mathrm{WD}}$ and $H_{\mathrm{VN}}^{\mathrm{SD}}$ are as follows:

$$J_{\mathrm{VN}}^{\mathrm{WD}} = \frac{1}{2|V|}\sum_{(u,v)\in E}\left\{\frac{\frac{d_u^{\mathrm{in}}}{d_u^{\mathrm{out}}} + \frac{d_v^{\mathrm{in}}}{d_v^{\mathrm{out}}}}{d_u^{\mathrm{out}} d_v^{\mathrm{in}}}\right\}, \tag{25}$$

$$J_{\mathrm{VN}}^{\mathrm{SD}} = \frac{1}{2|V|}\sum_{(u,v)\in E}\left\{\frac{d_u^{\mathrm{in}}}{d_v^{\mathrm{in}} d_u^{\mathrm{out}^2}}\right\}. \tag{26}$$

It is important to note that these normalized quantities and the original entropy measures have opposite monotonicity properties. In other words, when the normalized entropy J_{VN}^{D} decreases, the approximate von Neumann entropy H_{VN}^{D} increases.

A. Datasets used

We commence by giving a brief overview of the datasets used for experiments in this paper. We use four different datasets: the first two are synthetically generated artificial networks, while the other two are extracted from real-world systems.

Dataset 1. Consists of 10 directed networks evolved under preferential attachment. Each network starts from a fully connected seed network of five nodes. At each time step, a new node is added to the network. This node connects with nodes already in the network with a probability proportional to the steady state probability of a random walk taking place in the network. See Antiqueira *et al.* [23] for details about the model.

Dataset 2. Contains a large number of directed graphs which are randomly generated according to one of three different directed random graph models, namely, (a) the classical Erdős-Rényi model, (b) the "small-world" model, introduced by Watts and Strogatz [24], and (c) the "scale-free" model, developed by Barabási and Albert [25]. The different directed graphs in the database are created using a variety of model parameters, e.g., the graph size and the connection probability in the Erdős-Rényi model, the link rewiring probability [5] in the "small-world" model, and the number of added connections at each time step [5] in the "scale-free" model.

Dataset 3. It is extracted from the protein database previously used by Riesen and Bunke [26]. It consists of over 200 graphs, representing proteins from the Protein Data Bank [27], labeled with their corresponding enzyme class labels from the BRENDA enzyme database [28]. The database consists of six classes (labeled EC 1, . . . , EC 6), which represent proteins out of the six enzyme commission top level hierarchy (EC classes). The proteins are converted into graphs by first replacing the secondary structure elements of a protein with nodes, and then constructing a three-nearest-neighbor graph for the secondary structure elements. The graphs are thus directed.

Dataset 4. The arXiv HEP-TH (high energy physics theory) citation network. This is an evolving citation graph (Gehrke *et al.* [29]) extracted from the *e*-print arXiv. The directed network represents the citations within a data set of 27 770 papers by 352 807 directed links. If a paper u cites paper v, then the graph contains a directed link from node u to node v. Since there is no information about papers that are not included in the database, we do not consider such papers in the network. The data cover papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its HEP-TH section [30].

An important point to note concerning these datasets is that in Sec. II, to keep our development simple and straightforward, we require the directed graph under study be strongly connected, but here the graphs used for experiments do not always guarantee the strong connectivity, which implies that the graphs may have more than one strongly connected component. In fact, by simply summing up the entropy for each strongly connected component in a graph, our suggested approximate von Neumann entropy can also apply to directed graphs that are not strongly connected.

Moreover, it is also worth noting that in Dataset 4, citation networks do not contain bidirectional links (a paper can not cite any paper that has not yet been written). As a result, they are strongly directed graphs that contain a number of sink nodes. According to our previous analysis, these sink nodes may lead to situations where the directed graph von Neumann entropy is not well defined. However, from the strongly directed von Neumann entropy approximation obtained in Eq. (23), we find for a directed link $(u,v) \in E$ the denominator term $d_v^{\mathrm{in}} d_u^{\mathrm{out}^2}$ is only related to the out-degree of the starting node u and the

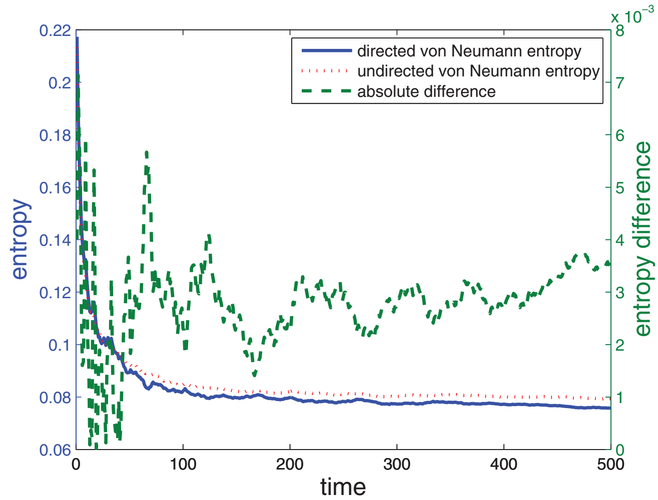


FIG. 1. (Color online) Comparing approximate von Neumann entropy for directed graphs J_{VN}^{D} [Eq. (24)] with approximate von Neumann entropy for undirected graphs J_{VN}^{U} [Eq. (27)]. Blue solid line: directed von Neumann entropy; red dotted line: undirected von Neumann entropy; green dashed line: absolute difference.

in-degree of the end node v. This means that the sink nodes do not make the expression singular, so the strongly directed von Neumann entropy approximation can still be computed in a valid manner on citation networks.

B. Von Neumann entropy of directed and undirected graphs

We first investigate the difference between the previously defined undirected graph von Neumann entropy and its directed analog in order to analyze how these entropies correlate. To do this, we select the directed graphs in Dataset 1 and compute their normalized entropies using Eq. (24), we then drop all the link directions to make the graphs undirected and compute their corresponding entropies using the following normalized quantity:

$$J_{\mathrm{VN}}^{U} = \frac{1}{|V|^2} \sum_{(u,v)\in E} \frac{1}{d_u d_v}. \tag{27}$$

Figure 1 shows the mean of the normalized entropies and their difference versus graph size for both directed and the corresponding undirected graphs. The main feature to note is that as time evolves, the difference between the two normalized entropies maintains small, which suggests that the directed and undirected graph von Neumann entropies have consistence on graphs that are well correlated. Moreover, it is clear from the plot that at some particular time, the directed entropy (blue solid line) fluctuates significantly while the corresponding undirected one (red dotted line) does not; as a result, the difference between them becomes particularly large. This implies that by dropping the link directions, the undirected graph obtained loses some of the structural information residing in the directed graph. Thus, the undirected graph von Neumann entropy also fails to capture this structure.

In Sec. II, we made use of the assumption that the local average node degree ratio r_u [Eq. (7)] is close to unity in order to develop our approximate expression for the von Neumann entropy. In order to explore whether this assumption is empirically valid, we explore the dependence of the approximate entropy on the average value of the node degree ratio. To this end, we compute the average of the local degree ratios over all nodes in a graph, i.e.,

$$r = \frac{1}{|V|} \sum_{u\in V} r_u, \tag{28}$$

where r_u is the degree ratio for node u. We investigate empirically how this global ratio affects the accuracy of the approximate von Neumann entropy.

We commence by studying some real-world directed networks and compare their normalized approximate von Neumann entropies [Eq. (24)] with normalized true entropies, which are computed using the formula

$$J_{\mathrm{TVN}}^{D} = |V| \left| H_{\mathrm{TVN}}^{D} - \left(1 - \frac{1}{|V|}\right) \right|. \tag{29}$$

The networks under study include the Wikipedia vote network, provided by Leskovec *et al.* [31], the Gnutella peer-to-peer networks from August 5 to 9, 2002, which are a sequence of snapshots of the Gnutella peer-to-peer file sharing network [32] and the arXiv HEP-TH citation network in Dataset 4. Table I gives the network size, the minimum and maximum values of the local degree ratio, and the average degree ratio. The table also lists the values of both the true entropy and approximate entropy. For each network studied, the average node degree ratio is always between 0 and 1, although locally the degree ratio differs significantly. The main feature to note is that the difference between the true and approximate entropy is relatively small, even though we are dealing with large networks.

To take this analysis a step further, we use the random directed graphs in Dataset 2. These graphs are generated according to three different models, and we use them to investigate the degree to which the approximate entropy deviates from the true value for different types of structure. We generate 1000 graphs for each model. For each graph,

TABLE I. Average node degree ratio and relative error for real-world network data.

Datasets	Wiki-Vote	p2p-G05	p2p-G06	p2p-G08	p2p-G09	ArXiv HEP-TH
Graph size	8297	8846	8717	6301	8114	27751
Min. ratio	0.0213	0.0303	0.0177	0.0208	0.0182	0.0035
Max. ratio	24.7500	22.4583	12.0000	9.2222	19.7000	46.6667
Average ratio	0.1984	0.4663	0.4501	0.4335	0.4196	0.4874
True entropy	0.0128	0.0343	0.0361	0.0374	0.0357	0.0430
Approx. entropy	0.0142	0.0489	0.0412	0.0387	0.0419	0.0585

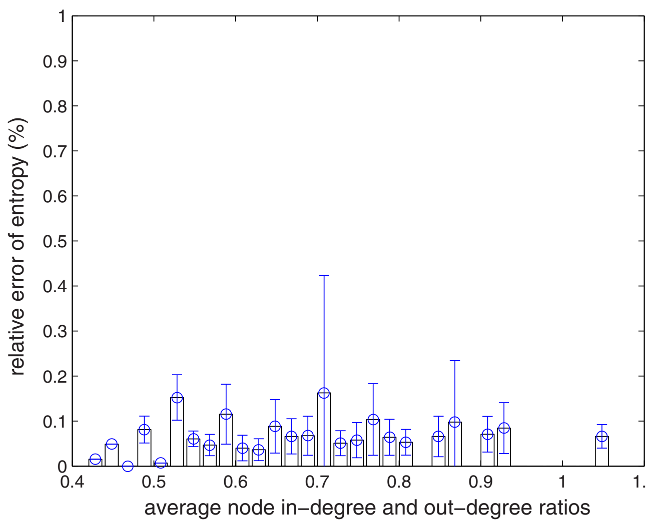


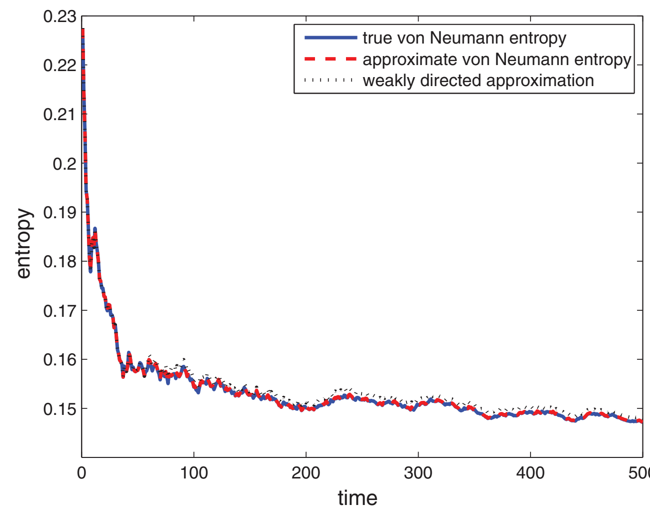
FIG. 2. (Color online) Mean and standard deviations of relative error of normalized approximate von Neumann entropy J_{VN}^{D} given in Eq. (24) for graphs with different average (global) node in-degree and out-degree ratios defined in Eq. (28).

we compute the relative error in the normalized approximate entropy, i.e., $\frac{|J_{\mathrm{VN}}^{D}-J_{\mathrm{TVN}}^{D}|}{J_{\mathrm{TVN}}^{D}}$. We then calculate the mean and standard deviations of the relative error, and explore the dependence on the global node degree ratio defined in Eq. (28). Figure 2 shows the mean and standard deviations (standard deviation shown as an error bar) of the relative error as a function of the global node degree ratio. The statistics needed for this plot are accumulated over graphs whose average node degree ratio falls into a fixed interval. From the plot it is clear the relative error is negligible (less than 0.2%) for graphs with global node degree ratios ranging between 0.4 and 1.1. Moreover, it takes on its minimum value when the ratio is equal to unity. This is as expected since our development of approximate von Neumann entropy expression is based on the assumption [Eq. (4)] that local nodes have the similar in-degree and out-degree. Therefore, the experimental results demonstrate that the approximate von Neumann entropy does not deviate too far from the true value even when the global node degree ratio is not close to unity and thus our assumption appears empirically valid.

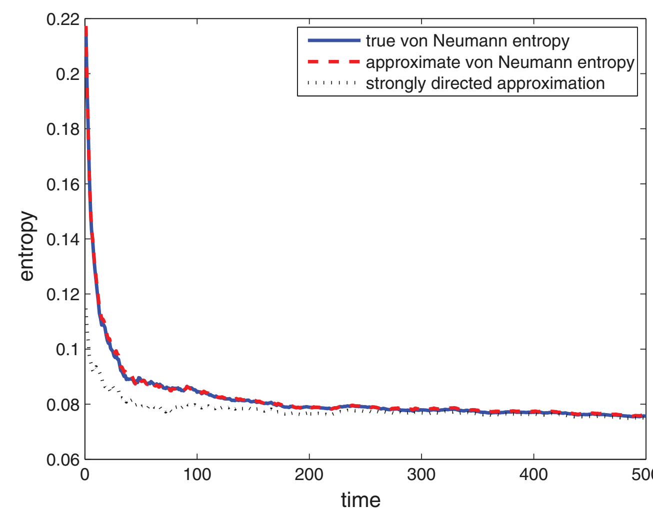
C. Weakly and strongly von Neumann entropy of directed graphs

In this section, we aim to use Dataset 1 to examine the accuracy of the approximations of the entropy for weakly and strongly directed graphs. In other words, we verify that the simplified expressions approximate well the true values of von Neumann entropy. In fact, the evolving directed graphs in Dataset 1 are strongly directed as the number of unidirectional links is significantly greater than that of bidirectional links. To obtain weakly directed graphs, we choose a large number of pairs of nodes that are connected by unidirectional links in these strongly directed graphs, and change the unidirectional connections to bidirectional ones.

In Figs. 3(a) and 3(b), we show the mean of the normalized entropies versus graph size for the directed graphs in Dataset 1.



(a)



(b)

FIG. 3. (Color online) Approximations to the von Neumann entropy: (a) comparing the weakly directed approximation $J_{\mathrm{VN}}^{\mathrm{WD}}$ given in Eq. (25) to the values of J_{TVN}^{D} in Eq. (29) and J_{VN}^{D} in Eq. (24); (b) comparing the strongly directed approximation $J_{\mathrm{VN}}^{\mathrm{SD}}$ given in Eq. (26) to the values of J_{TVN}^{D} in Eq. (29) and J_{VN}^{D} in Eq. (24). Blue solid line: true von Neumann entropy; red dashed line: approximate von Neumann entropy; black dotted line: weakly and strongly directed approximations.

Here, we have computed the approximate entropies for weakly and strongly directed graphs $J_{\mathrm{VN}}^{\mathrm{WD}}$ and $J_{\mathrm{VN}}^{\mathrm{SD}}$ using Eqs. (25) and (26), respectively. We compare their values with the normalized approximate entropy J_{VN}^{D} given in Eq. (24) and the normalized true entropy J_{TVN}^{D} defined in Eq. (29).

From both plots, as the network evolves, all these quantities decrease gradually to a value close to zero, which implies that the true von Neumann entropy and its approximations increase monotonically until a plateau value of unity is reached. It is also worth noting that the difference between these entropies is negligible, thus we can deduce that the approximate von Neumann entropy we suggested (red dashed line) approximates the true von Neumann entropy (blue solid line) very well.

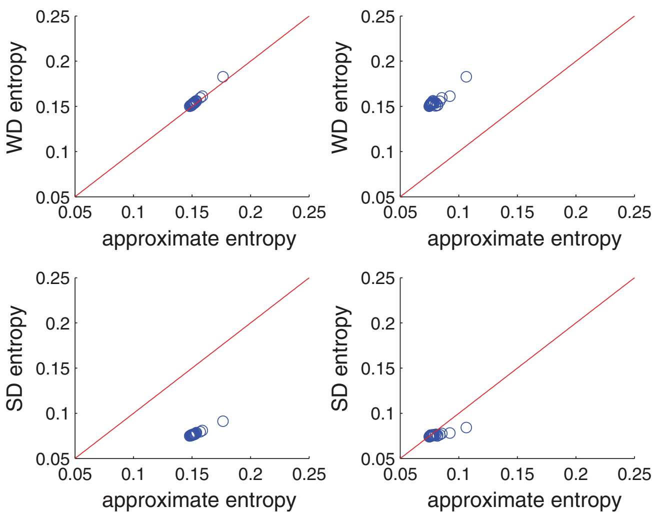


FIG. 4. (Color online) Comparing approximate entropies for weakly directed (WD) graphs $J_{\mathrm{VN}}^{\mathrm{WD}}$ [Eq. (25)] and strongly directed (SD) graphs $J_{\mathrm{VN}}^{\mathrm{SD}}$ [Eq. (26)] with approximate entropy J_{VN}^{D} [Eq. (24)]. Top left: $J_{\mathrm{VN}}^{\mathrm{WD}}$ vs J_{VN}^{D} for weakly directed graphs; top-right: $J_{\mathrm{VN}}^{\mathrm{WD}}$ vs J_{VN}^{D} for strongly directed graphs; bottom left: $J_{\mathrm{VN}}^{\mathrm{SD}}$ vs J_{VN}^{D} for weakly directed graphs; bottom right: $J_{\mathrm{VN}}^{\mathrm{SD}}$ vs J_{VN}^{D} for strongly directed graphs.

Figure 4 shows scatter plots of the weakly and strongly directed approximations $J_{\mathrm{VN}}^{\mathrm{WD}}$ and $J_{\mathrm{VN}}^{\mathrm{SD}}$ versus the approximate entropy J_{VN}^{D} for sets of weakly directed and strongly directed graphs. We select the relevant sets of graphs from Dataset 1 using a fixed time interval, which gives 50 samples for strongly and weakly directed graphs respectively.

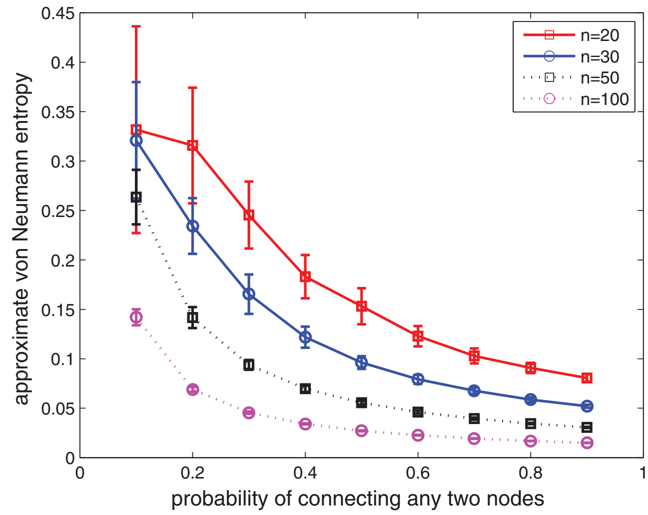
It is clear from Fig. 4 that the scatter plots of the weakly (strongly) directed approximations $J_{\mathrm{VN}}^{\mathrm{WD}}$ ($J_{\mathrm{VN}}^{\mathrm{SD}}$) are much closer to the true values for the weakly (strongly) directed graphs J_{VN}^{D}. Thus, we conclude that the true value of von Neumann entropy and the simplified weakly (strongly) directed form we suggested are approximately equivalent on weakly (strongly) directed graphs.

D. Von Neumann entropy for distinguishing directed graphs

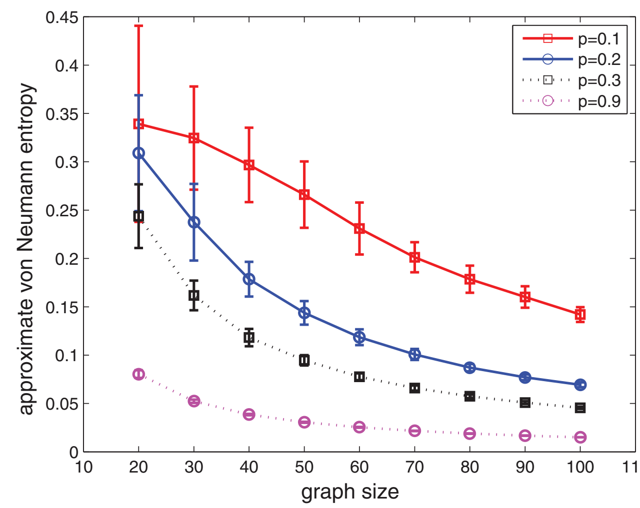
We aim to explore whether the von Neumann entropy can be used to distinguish directed graphs with different structural properties. To this end, we have generated graphs with different parameter settings and explored the dependence of the von Neumann entropy on these parameters.

We commence by considering the Erdős-Rényi model, where the two parameters are the graph size n (or number of nodes) and the node link probability p. We vary these parameters and randomly generate a number of directed graphs at each setting. We compute the mean and standard deviations of the normalized approximate von Neumann entropy J_{VN}^{D} [Eq. (24)] over samples with the same parameter settings.

Figure 5(a) shows the normalized approximate von Neumann entropy (mean and standard deviations as an error bar) for the Erdős-Rényi model, with $n = 20,30,50,100$ as a function of p varying from 0.1 to 0.9. Figure 5(b) plots the same data for $p = 0.1,0.2,0.3,0.9$ as a function of n varying from 20 to 100. From the plots it is clear that the mean value of the normalized entropy decreases gradually, which implies that



(a)



(b)

FIG. 5. (Color online) Mean and standard deviations of the normalized approximate von Neumann entropy J_{VN}^{D} computed using Eq. (24) as a function of (a) node link probability p and (b) graph size n for Erdős-Rényi graphs. Red square solid line: (a) $n = 20$, (b) $p = 0.1$; blue circle solid line: (a) $n = 30$, (b) $p = 0.2$; black square dotted line: (a) $n = 50$, (b) $p = 0.3$; magenta circle dotted line: (a) $n = 100$, (b) $p = 0.9$.

the von Neumann entropy increases with both the graph size and the node link probability. This result is as expected since in an Erdős-Rényi network, the structure becomes more complex when there are both a large number of nodes in the network (n is large) and there are a large number of random links in the network (p is large). When the probability p is small, the standard deviation of entropy is particularly large. This is because for a network with a fixed size, a smaller number of directed links in the network leads to a greater uncertainty of how these links are connected. As a result, there is significant variance in the network entropy.

To take this analysis a step further, we apply the von Neumann entropy to the directed graphs in the artificial network data (Dataset 2) to investigate whether different topologies can be distinguished.

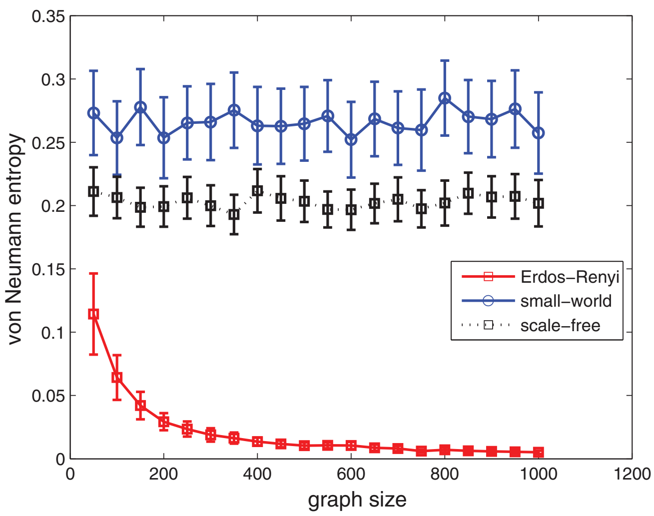


FIG. 6. (Color online) Mean and standard deviations of the approximate von Neumann entropy quantity J_{VN}^{D} [Eq. (24)] for different models of directed graphs versus graph size. Red square solid line: Erdős-Rényi; blue circle solid line: "small-world"; black square dotted line: "scale-free."

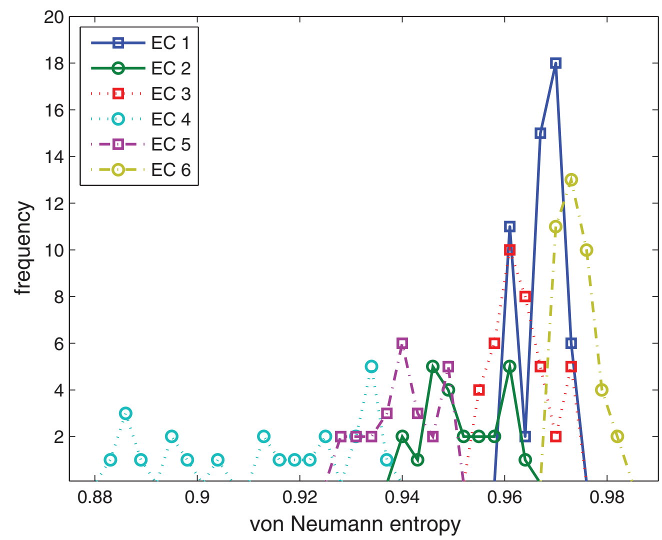


FIG. 7. (Color online) Histograms of von Neumann entropy H_{VN}^{D} [Eq. (21)] for different enzyme classes. Blue square solid line: EC 1; green circle solid line: EC 2; red square dotted line: EC 3; cyan circle dotted line: EC 4; purple square dash-dot line: EC 5; Khaki circle dashed-dotted line: EC 6.

Figure 6 shows the mean value of the normalized approximate von Neumann entropy as a function of graph size (again, standard deviation as an error bar). For a given graph size, the difference in mean entropy for different models is much larger than the standard deviation of the entropy within each model. This suggests that the variance in von Neumann entropy due to different parameter settings is much smaller than that due to differences in structure, which means that different network models have different values of von Neumann entropy for a given size.

Finally, we aim to verify whether the von Neumann entropy can be used to determine the enzyme class of the protein graphs. Here, in order to obtain a more visual-friendly plot, we use the original approximate von Neumann entropy expression (21) instead of the normalized entropy (24). In Fig. 7, we show a histogram of the von Neumann entropy for the graphs in the database. The different line styles represent different enzyme classes in the database. Four classes of proteins (EC 3, EC 4, EC 5, and EC 6) show some separation. Another interesting feature is that class EC 1 is located between and is also overlapped with classes EC 3 and EC 6. Because of the larger population of EC 1, the overlap is in fact relatively small. Unfortunately, class EC 2, on the other hand, can not be easily separated as it is mixed with classes EC 3 and EC 5.

The experiments in this section show that the directed von Neumann entropy can be efficiently used to distinguish different types of directed graphs from both artificial and real-world data since it captures differences in structural features of directed networks.

E. Von Neumann entropy for analyzing citation networks

Next, we explore whether the von Neumann entropy and its simplifications for strongly and weakly directed graphs can be used to detect changes in the structure of a citation network that evolves over time. In this context, it is important to note that a high impact (or highly cited) paper may cause a much more significant change in the network structure than a paper with an average citation profile since such a paper usually represents a paradigm change in the subject it concerns.

We convert the arXiv HEP-TH citation network to an evolving directed graph and explore whether the directed von Neumann entropy can be used to detect changes in graph structure caused by the publication of high impact papers.

As noted earlier, Dataset 4 is hermetic in the sense that it does not contain any citation information related to papers that fall outside its coverage. Thus, the citation graph grows from a single node to a graph consisting of 27 770 nodes with 352 807 directed links. Occasionally, a newly published paper may cite a number of papers that are not in the current citation network, i.e., these papers do not cite any papers in the dataset and are only cited by other papers in it. In this case, we regard the newly published paper as a primary paper and the cited papers as its secondary papers. The primary paper and the secondary papers are thus introduced into the network at the same time epoch.

There are 25 001 primary papers in the dataset, and we label them from 1 to 25 001 according to the time at which they first appear in the citation network. Hence, these ordinal labels index the epoch at which papers appear in the database and can be viewed as a time sequence, i.e., the citation network begins at $t = 1$ (January 1993) and ends at $t = 25\,001$ (April 2003).

The impact of a paper on a citation network is not reflected immediately after it is published. Instead, the influence develops and is sustained for a period of time. This is because after the publication of a high impact paper, a large number of subsequent papers will cite it (in the citation graph, the corresponding node will have a large in-degree). As a result, its influence will be sustained until the most recent paper has cited it. In order to capture the impact of a paper, we use the rate of change of the directed von Neumann entropy.

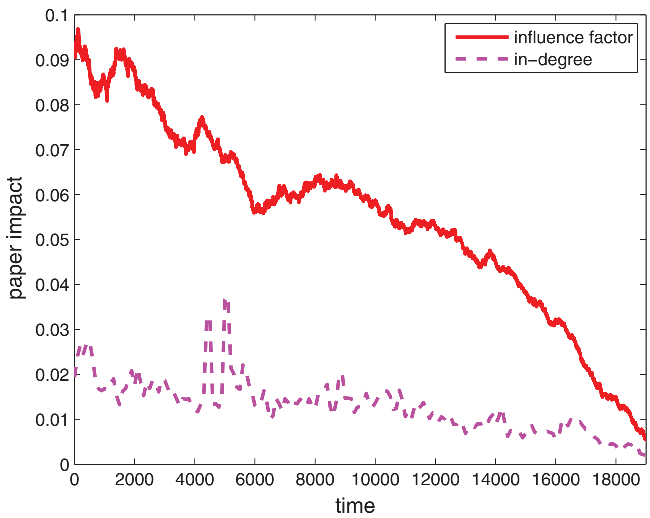


FIG. 8. (Color online) Citational influence factor σ_u in Eq. (30) and average node in-degree versus time for the arXiv HEP-TH citation network. Red solid line: influence factor; magenta dashed line: average node in-degree.

To use this quantity to measure the citational influence of a paper, suppose a primary paper u is published at time t_0, and its impact is sustained for a period of length N, which means the impact ends at time t_N. We define the citational influence factor σ_u of paper u as the mean value of the relative change in the normalized strongly directed approximation $J_{\rm VN}^{\rm SD}$ [Eq. (26)] over the relevant influence period $t_0, t_1, \ldots, t_N$, i.e.,

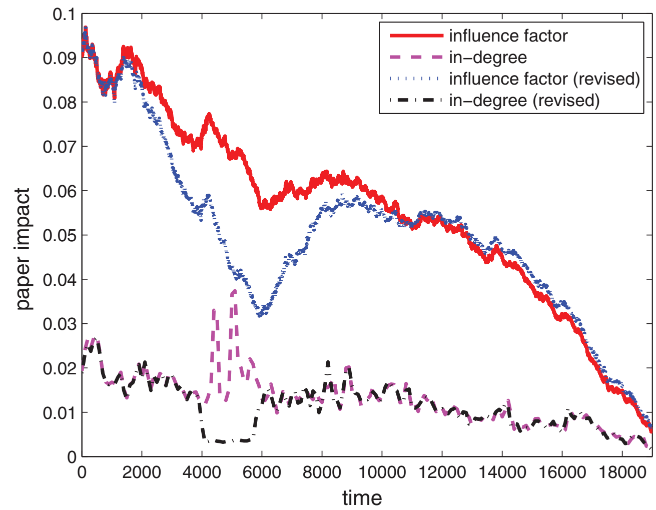
$$\sigma_u \triangleq \frac{\sum_{i=1}^{N}\left(J_{\rm VN}^{\rm SD}(t_i) - J_{\rm VN}^{\rm SD}(t_0)\right)}{N\, J_{\rm VN}^{\rm SD}(t_0)}. \tag{30}$$

From the dataset, we find that most papers have an influence period between 5000 to 6000 (measured in terms of change in sequence number). Thus, we take the average and fix $N = 5500$. At the beginning of the citation sequence, the volume of data is not sufficient for reliable analysis. We thus start the analysis at $t = 5000$ instead of $t = 1$ and terminate at $t = 24\,000$ which gives a sequence length of 19 000.

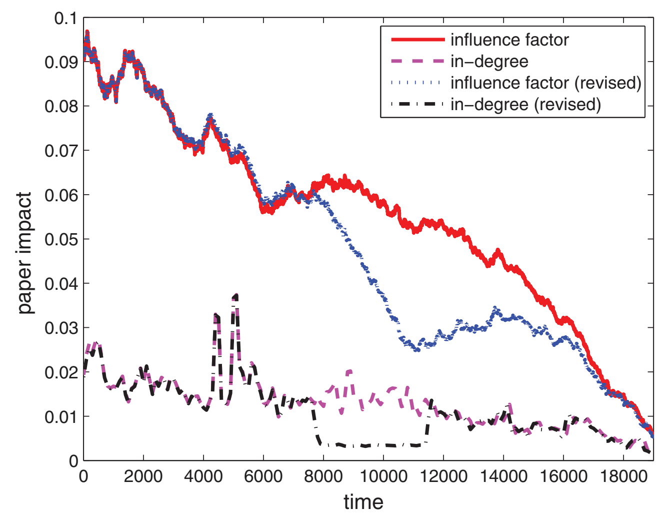
In Fig. 8, we plot both the influence factor and the in-degree distribution for primary papers against time. The main feature to note is that although the influence factor fluctuates, it decreases gradually to a value close to zero. This is because as time evolves, the citation network size increases rapidly, reducing the potential relative impact of more recent papers.

Another important feature of this figure is that our influence factor can be used to reveal the changes in structure caused by influential papers. In the plot at epochs close to $t = 2000, 4500, 14\,000$, we see some significant fluctuations in the influence factor curve, which represent significant changes in network structure. Turning our attention to the in-degree distribution, there are peaks at epochs around $t = 2000, 4500, 14\,000$, which means that papers published at these times are cited heavily. Thus, we combine these observations and suggest that the influential papers can create significant changes in the structure of the evolving citation network.

To take this analysis a step further, we modify the original citation data and explore how the influence factors change. To this end, we select papers from a period of time and



(a)



(b)

FIG. 9. (Color online) Paper impact measures versus time for modified citation networks when citation data are deleted (a) between 4000 and 6000 and (b) between 8000 and 12 000. Red solid line: original citational influence factor; magenta dashed line: original average node in-degree; blue dotted line: citational influence factor after modification; black dashed-dotted line: average node in-degree after modification.

delete most of their citation connections. Figure 9 shows the analysis before and after modifying the data if we delete connections in the time interval (a) $t \in [4000, 6000]$ and (b) $t \in [8000, 12\,000]$. As a result, the revised influence factors show a sharp drop in values, but after a transient time return to the behavior of the original curve. Thus, there are significant differences in the network structure when high impact papers are published, and the directed graph von Neumann entropy can capture such differences.

IV. CONCLUSIONS

This paper is motivated by the aim of developing a an effective entropy measure for quantifying the structural complexity of directed graphs. We have made a number of contributions. First, we have shown how to compute the von Neumann entropy of a directed graph using Chung's definition of the normalized Laplacian matrix. We simplify the calculation of von Neumann entropy by replacing the Shannon entropy by the quadratic entropy. From this starting point, we have developed approximations to the entropy that can be computed using in-degree and out-degree statistics. Moreover, we present specific approximations of the von Neumann entropy that apply to both strongly and weakly directed graphs, according to whether or not the majority of links are unidirectional links.

To evaluate these methods and analyze their properties, we have undertaken experiments on both artificial and real-world network data. These experiments demonstrate that the suggested von Neumann entropy for directed graphs can be used to characterize different classes of proteins and analyze the structural properties of an evolving citation network. Moreover, we show that the entropy characterization is not unduly limited by the approximations made in deriving it.

The work reported in this paper can clearly be extended in a number of different ways. First, we acknowledge that we have explored a relatively limited quantity of real-world data. It would, for example, be interesting to see if the method can be used to detect network anomalies and disturbances. Another interesting line of investigation would be to explore whether the method can be used to analyze page rank matrices since these too are based on random walks on directed graphs. Finally, we plan to explore whether this work can be extended to edge-weighted graphs, labeled graphs, and hypergraphs.

ACKNOWLEDGMENTS

The authors are grateful to FAPESP Project York-USP (Project No. 12/50986-7) for financial support. C. H. Comin acknowledges FAPESP (Grant No. 11/22639-8) for financial support. L. d. F. Costa thanks FAPESP (Grant No. 11/50761-2) and CNPq (Grant No. 573583/2008-0) for financial support. E. R. Hancock acknowledges support through a Royal Society Wolfson Research Merit Award.

[1] M. Newman, SIAM Rev. **45**, 167 (2003).
[2] R. Albert and A. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[3] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[4] L. d. F. Costa, O. Oliveira, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. Correa Rocha, Adv. Phys. **60**, 329 (2011).
[5] E. Estrada, *The Structure of Complex Networks: Theory and Applications* (Oxford University Press, Oxford, UK, 2011).
[6] L. d. F. Costa and F. A. Rodrigues, Europhys. Lett. **85**, 48001 (2009).
[7] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, UK, 2010).
[8] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, Adv. Phys. **56**, 167 (2007).
[9] F. Passerini and S. Severini, Int. J. Agent Technol. Syst. **1**, 58 (2009).
[10] F. Chung, Ann. Combin. **9**, 1 (2005).
[11] E. Estrada, M. Fox, D. Higham, and G. Oppo (Eds.) *Network Science. Complexity in Nature and Technology* (Springer, Berlin, 2010), XI, p. 245.
[12] D. Feldman and J. Crutchfield, Phys. Lett. A **238**, 244 (1998).
[13] M. Dehmer, A. Mowshowitz, and F. Emmert-Streib, *Advances in Network Complexity* (Wiley-Blackwell, Boston, 2013).
[14] S. Riis, Comput. Res. Repository. **0711** (2007).
[15] D. Berwanger, E. Gradel, L. Kaiser, and R. Rabinovich, Theor. Comput. Sci. **463**, 2 (2012).
[16] F. Escolano, B. Bonev, and E. Hancock, *Structural, Syntactic, and Statistical Pattern Recognition* (Springer, Berlin, 2012), pp. 190–198.
[17] F. Escolano, E. Hancock, and M. Lozano, Phys. Rev. E **85**, 036206 (2012).
[18] K. Anand, G. Bianconi, and S. Severini, Phys. Rev. E **83**, 036109 (2011).
[19] S. Braunstein, S. Ghosh, and S. Severini, Ann. Combin. **10**, 291 (2006).
[20] N. Biggs, *Algebraic Graph Theory. Cambridge Tracts in Mathematics No. 67* (Cambridge University Press, Cambridge, 1974).
[21] L. Han, F. Escolano, E. Hancock, and R. Wilson, Pattern Recog. Lett. **33**, 1958 (2012).
[22] S. Ihara, *Information Theory for Continuous Systems* (World Scientific Pub. Co. Inc., Singapore, 1993), p. 308.
[23] L. Antiqueira, F. Rodrigues, and L. d. F. Costa, J. Stat. Mech.: Theor. Exper. (2009) L09005.
[24] D. Watts and S. Strogatz, Nature (London) **393**, 440 (1998).
[25] A. Barabási and R. Albert, Science **286**, 509 (1999).
[26] K. Riesen and H. Bunke, in *Structural, Syntactic, and Statistical Pattern Recognition* (Springer, Berlin, 2008), Vol. 5342, pp. 287–297.
[27] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shidyalov, and P. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[28] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, Nucleic Acids Res. **32**, 081 (2004).
[29] J. Gehrke, P. Ginsparg, and J. Kleinberg, SIGKDD Explor. **5**, 149 (2003).
[30] J. Leskovec, J. Kleinberg, and C. Faloutsos, in *Proceedings ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD)* (ACM, New York, 2005), pp. 177–187.
[31] J. Leskovec, D. Huttenlocher, and J. Kleinberg, in *Proceedings ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)* (ACM, New York, 2010), pp. 1361–1370.
[32] M. Ripeanu, I. Foster, and A. Iamnitchi, IEEE Internet Computing **6**, 50 (2002).