Entropy and Heterogeneity Measures 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, UK
{cy666,richard.wilson,edwin.hancock}@york.ac.uk

[2] Institute of Physics at São Carlos, University of São Paulo, P.O. Box 369, São Carlos, São Paulo, 13560-970, Brazil
{appdnails,ldfcosta}@gmail.com

Abstract. In this paper, we aim to develop novel methods for measuring the structural complexity for 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 a number of alternative techniques that are applicable to directed graphs. We commence by using Chung's generalisation 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 vertex 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. Next we explore how to extend Estrada's heterogeneity index from undirected to directed graphs. Our measure is motivated by the simplified von Neumann entropy, and involves measuring the heterogeneity of differences in in-degrees and out-degrees. Finally, we perform an analysis which reveals a novel linear relationship between heterogeneity and resistance distance (commute time) statistics for undirected graphs. This means that the larger the difference between the average commute time and shortest return path length between pairs of vertices, the greater the heterogeneity index. Based on this observation together with the definition of commute time on a directed graph, we define an analogous heterogeneity measure for directed graphs. We illustrate the usefulness of the measures defined in this paper for datasets describing Erdos-Renyi, 'small-world', 'scale-free' graphs, Protein-Protein Interaction (PPI) networks and evolving networks.

Keywords: directed graph, structural complexity, von Neumann entropy, heterogeneity index.

⋆ Edwin R. Hancock is supported by a Royal Society Wolfson Research Merit Award.

E. Hancock and M. Pelillo (Eds.): SIMBAD 2013, LNCS 7953, pp. 219–234, 2013.

1 Introduction

Recently there has been considerable interest in analyzing the properties of complex networks since they play a significant role in modelling large-scale systems in biology, physics and the social sciences. In fact, complex networks provide convenient models for complex systems. However, to render such models tractable, it is essential to have to hand methods for characterizing their salient properties. As Costa and Rodrigues [7] stated, complex networks are graphs whose connectivity properties deviate from those of regular graphs, which can be defined as a process of being 'simple', and the complexity of a network can be understood as the complement of simplicity. Structural complexity is therefore perhaps the most important property of a complex network. In order to analyze the features of a complex network it is imperative that computationally efficient measures are to hand that can be used to represent and quantify the structural complexity.

In this context graph theoretic methods are often used since they provide effective tools for characterizing network structure together with the intrinsic complexity. This approach has lead to the design of several practical methods for characterizing the global and local structure of undirected networks. However, there is relatively little work aimed at characterizing directed network structure. One of the reasons for this is that the graph theory underpinning directed networks is less developed than that for undirected networks.

The aim in this paper is to explore whether a number of different characterizations developed for undirected graphs can be extended to the domain of directed graphs, using some recent results from spectral graph theory.

1.1 Related Literature

Recently, Amancio et al. [1] have shown that labyrinths can be modelled as complex networks and studied in terms of the concept of absorption time, defined as the time it takes for a random walk from an internal node to an output node, to classify networks' metrics. Moreover, Estrada [10] has proposed an index that can be used to quantify the heterogeneity characteristics of undirected graphs. This index depends on vertex degree statistics and graph size. The lower bound of this quantity is zero, which occurs for a regular graph (i.e. all the vertices have the same degree). The upper bound is equal to one, which is obtained for a star graph (i.e. there exists a central vertex and all other vertices connect and only connect to it).

Working in the domain of structural pattern recognition, Xiao et al. [19] have explored how the heat kernel trace can be used as a means to characterize the structural complexity of graphs. To do this, they first consider the zeta function associated with the Laplacian eigenvalues and use the derivative of zeta function at origin as a characterization for distinguishing different types of graphs. Ren et al. [15] have developed a novel method to characterize unweighted graphs by using the polynomial coefficients determined by the Ihara zeta function. To do this, they construct a pattern vector of Ihara coefficients, and successfully use this to cluster unweighted graphs. Furthermore, they extend their work by applying

Ihara coefficients from unweighted graphs to edge-weighted graphs, which is achieved by establishing the Perron-Frobenius operator with the assistance of a reduced Bartholdi zeta function.

Escolano et al. [8] have used the concept of thermodynamic depth to measure the complexity of networks. They first define the polytopal complexity of a graph and then introduce a phase-transition principle which links this complexity to the heat flow, and thus obtain a complexity measure referred as flow complexity. Recently, Han et al. [12] have developed simplified expressions of von Neumann entropy on undirected graphs. To do this, they replace the Shannon entropy by its quadratic counterpart, investigate how to simplify and approximate the calculation of von Neumann entropy. They also explore the relationship among the heterogeneity index, commute time and the von Neumann entropy, and introduce a graph complexity measure based on thermodynamic depth.

The above provides a brief survey of recent work on the structural complexity of undirected graphs. However, in the real world, directed graphs are also common as many networks can be modelled with them. For example, the World Wide Web is a directed network in which vertices represent web pages while edges are the hyperlinks between pages.

Turning our attention to directed graphs, Riis [16] has extended the concept of entropy to directed graphs, using the definitions 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 size. Berwanger et al. [4] have proposed a new parameter for the complexity of infinite directed graphs by measuring to what extent the cycles in graphs are intertwined. This index is defined according to the definitions of tree width, directed tree width and hypertree width and a similar 'robber-and-cops' game. Recently Escolano et al. [9] have extended the concept of heat diffusion thermodynamic depth for undirected networks to directed networks and thus obtain a measure to quantify the complexity of structural patterns encoded by directed graphs.

1.2 Paper Outline

One natural way of capturing the structure of directed networks is to use statistics that capture the balance of in-degree and out-degree at vertices. In this paper we commence from Passerini and Severini's work [13], 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 [6] definition of the normalized Laplacian on a directed graph. According to this definition, the directed normalized Laplacian is Hermitian, so Passerini and Severini's interpretation still holds for the domain of directed graphs. The von Neumann entropy is essentially the Shannon entropy associated with the normalized Laplacian eigenvalues. If we approximate the Shannon entropy by its quadratic counterpart, then the von Neumann entropy can be simplified. The resulting expression depends on the in-degree and out-degree of pairs of vertices connected by edges.

To simplify this resulting expression a step further, we consider graphs that are either weakly or strongly directed, i.e. those in which there are large or small proportions of bidirectional edges, and develop corresponding approximations of the von Neumann entropy.

Finally, we explore how Estrada's heterogeneity index can be extended from undirected to directed graphs. Our study of von Neumann entropy suggests a statistic determined by the in-degree and out-degree for nodes connected by a directed edge. We show that the resulting heterogeneity index is linked to the difference between the elements of the normalized adjacency matrix (as a local measure of connectivity) and the average commute time between nodes (or resistance distance) as a more global measure of connectivity structure.

The outline of this paper is as follows. In Sect.2, we develop the simplified forms of von Neumann entropy of directed graphs, and in Sect.3, we introduce the heterogeneity index and commute time on directed graphs and then investigate their correlation. In Sect.4, we analyze our theoretical result by undertaking experiment on network datasets and finally we conclude this paper with an evaluation of our contribution and suggestions for future work.

2 Von Neumann Entropy of Directed Graphs

In this section, we propose novel methods for characterizing the complexity of directed graphs. The first 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 vertices. We then provide approximations for the von Neumann entropy for both strongly directed graphs where there are few bidirectional edges and weakly directed graphs where there are few edges that are not bidirectional.

2.1 Laplacian of Directed Graphs

Suppose $G(V, E)$ is a directed graph with vertex set V and edge set $E \subseteq V \times V$, then the structure of this graph can be represented by a $|V| \times |V|$ adjacency matrix A as follows (where $|V|$ is the number of vertices in the graph)

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The in-degree and out-degree of vertex i are

$$d_i^{in} = \sum_{j=1}^{|V|} A_{ji}, \quad d_i^{out} = \sum_{j=1}^{|V|} A_{ij}. \tag{2}$$

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

$$P_{ij} = \begin{cases} \frac{A_{ij}}{d_i^{out}} & \text{if } (i,j) \in E \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

According to the Perron-Frobenius Theorem, for a strongly connected directed graph, the transition matrix P has a unique left eigenvector ϕ with $\phi(i) > 0, \forall i$ 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 walk is at vertex i is the sum of all incoming probabilities of vertices j satisfying $(j,i) \in E$, i.e. $\phi(i) = \sum_{j,(j,i)\in E} \phi(j)P_{ji}$, then we can obtain the following approximate equation

$$\frac{\phi(i)}{\phi(j)} \approx \frac{d_i^{in}}{d_j^{in}}. \quad (4)$$

As stated in Chung [6], if we let $\Phi = diag(\phi(1), \phi(2), ...)$, then the normalized Laplacian matrix of a directed graph can be defined as

$$\tilde{L} = I - \frac{\Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^T \Phi^{1/2}}{2}. \quad (5)$$

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

2.2 Von Neumann Entropy of Undirected Graphs

Having defined the prerequisites, we now show how the concept of von Neumann entropy can be extended from undirected to directed graphs. Passerini and Severini [13] have argued that the normalized Laplacian can be interpreted as the density matrix of an undirected graph, and hence the associated von Neumann entropy of the graph is the Shannon entropy associated with the normalized Laplacian eigenvalues, i.e.

$$H_{VN}^U = -\sum_{i=1}^{|V|} \frac{\tilde{\lambda}_i}{|V|} \ln \frac{\tilde{\lambda}_i}{|V|} \quad (6)$$

where $\tilde{\lambda}_i$, $i = 1, ..., |V|$ are the eigenvalues of the normalized Laplacian matrix.

Commencing from their definition, Han et al. [12] have shown that for an undirected graph $G(V,E)$, the Shannon entropy $H_S^U = -\sum_{i=1}^{|V|} \frac{\tilde{\lambda}_i}{|V|} \ln \frac{\tilde{\lambda}_i}{|V|}$ can be approximated by the quadratic entropy $H_Q^U = \sum_{i=1}^{|V|} \frac{\tilde{\lambda}_i}{|V|}(1 - \frac{\tilde{\lambda}_i}{|V|})$. As a result the von Neumann entropy of undirected graphs can be approximated by

$$H_{VN}^{U} = \frac{Tr[\tilde{L}]}{|V|} - \frac{Tr[\tilde{L}^2]}{|V|^2}. \tag{7}$$

For undirected graphs, the traces appearing in the above expression can be approximated by degree statistics, with the result that

$$H_{VN}^{U} = 1 - \frac{1}{|V|} - \frac{1}{|V|^2} \sum_{(i,j)\in E} \frac{1}{d_i d_j}. \tag{8}$$

2.3 Von Neumann Entropy of Directed Graphs

To extend the analysis of Han et al. [12] to directed graphs, we commence from (7) and 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 vertices. To commence, we turn to Chung's expression for the normalized Laplacian of directed graphs and write

$$\begin{aligned} Tr[\tilde{L}] &= Tr[I - \frac{\Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^T \Phi^{1/2}}{2}] \\ &= Tr[I] - \frac{1}{2} Tr[\Phi^{1/2} P \Phi^{-1/2}] - \frac{1}{2} Tr[\Phi^{-1/2} P^T \Phi^{1/2}]. \end{aligned} \tag{9}$$

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

$$\begin{aligned} Tr[\tilde{L}] &= Tr[I] - \frac{1}{2} Tr[P \Phi^{-1/2} \Phi^{1/2}] - \frac{1}{2} Tr[P^T \Phi^{1/2} \Phi^{-1/2}] \\ &= Tr[I] - \frac{1}{2} Tr[P] - \frac{1}{2} Tr[P^T]. \end{aligned} \tag{10}$$

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

$$Tr[\tilde{L}] = Tr[I] = |V|, \tag{11}$$

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

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

$$\begin{aligned} Tr[\tilde{L}^2] &= Tr[I^2 - (\Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^T \Phi^{1/2}) + \\ &\quad \frac{1}{4}(\Phi^{1/2} P \Phi^{-1/2} \Phi^{1/2} P \Phi^{-1/2} + \Phi^{1/2} P \Phi^{-1/2} \Phi^{-1/2} P^T \Phi^{1/2} + \\ &\quad \Phi^{-1/2} P^T \Phi^{1/2} \Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^T \Phi^{1/2} \Phi^{-1/2} P^T \Phi^{1/2})] \\ &= Tr[I^2] - Tr[P] - Tr[P^T] + \frac{1}{4}(Tr[P^2] + Tr[P \Phi^{-1} P^T \Phi] + Tr[P^T \Phi P \Phi^{-1}] + Tr[P^{T^2}]) \\ &= |V| + \frac{1}{2}(Tr[P^2] + Tr[P \Phi^{-1} P^T \Phi]), \end{aligned} \tag{12}$$

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

To continue the development we first divide the edge set E into two disjoint subsets E_1 and E_2, where $E_1 = \{(i,j)|(i,j) \in E \text{ and } (j,i) \notin E\}$, $E_2 = \{(i,j)|(i,j) \in E \text{ and } (j,i) \in E\}$ 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

$$Tr[P^2] = \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} P_{ij} P_{ji} = \sum_{(i,j)\in E_2} \frac{1}{d_i^{out} d_j^{out}}. \tag{13}$$

Using the fact that $\Phi = diag(\phi(1), (2), ...)$ we have

$$Tr[P\Phi^{-1}P^T\Phi] = \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} P_{ij}^2 \frac{\phi(i)}{\phi(j)} = \sum_{(i,j)\in E} \frac{\phi(i)}{\phi(j) d_i^{out^2}}. \tag{14}$$

Using (4), i.e. $\frac{\phi(i)}{\phi(j)} \approx \frac{d_i^{in}}{d_j^{in}}$, we can approximate the von Neumann entropy of a directed graph in terms of the in-degree and out-degree of the vertices as follows

$$H_{VN}^D = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2} \left\{ \sum_{(i,j)\in E} \left(\frac{1}{d_i^{out} d_j^{out}} + \frac{d_i^{in}}{d_j^{in} d_i^{out^2}} \right) - \sum_{(i,j)\in E_1} \frac{1}{d_i^{out} d_j^{out}} \right\}, \tag{15}$$

or equivalently,

$$H_{VN}^D = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2} \left\{ \sum_{(i,j)\in E} \frac{d_i^{in}}{d_j^{in} d_i^{out^2}} + \sum_{(i,j)\in E_2} \frac{1}{d_i^{out} d_j^{out}} \right\}. \tag{16}$$

We can simplify this expression a step further according to the relative sizes of the sets E_1 and E_2.

For weakly directed graphs, $|E_1| \ll |E_2|$, i.e. few of the edges are not bidirectional, and we can ignore the summation over E_1 in (15). Re-writing the remaining terms in curly braces in terms of a common denominator and then dividing numerator and denominator by $d_i^{out} d_j^{out}$ we obtain

$$H_{VN}^{WD} = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2} \sum_{(i,j)\in E} \frac{\frac{d_i^{in}}{d_i^{out}} + \frac{d_j^{in}}{d_j^{out}}}{d_i^{out} d_j^{in}}. \tag{17}$$

The first term $1 - \frac{1}{|V|}$ tends to unity as the graph size becomes large and the remaining term is normalized by $2|V|^2$. In its second term above, the numerator is given in terms of the sum of the ratios of in-degree and out-degree at the two vertices. Since the directed edges cannot commence at a sink (a node of zero out-degree), the ratios do not become infinite. Replacing $d_i^{out^2}$ in the denominator by $d_i^{in} d_i^{out}$, we obtain the following expression that approximates the von Neumann entropy for weakly directed graphs

$$H_{VN}^{WD} = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2} \sum_{(i,j)\in E} \left\{ \frac{1}{d_i^{out} d_j^{in}} + \frac{1}{d_i^{in} d_j^{out}} \right\}. \tag{18}$$

On the other hand, for strongly directed graphs, there are few bidirectional edges, i.e. $|E_2| \ll |E_1|$, and we can ignore the summation over E_2 in (16), giving the approximate entropy for strongly directed graphs

$$H_{VN}^{SD} = 1 - \frac{1}{|V|} - \frac{1}{2|V|^2} \sum_{(i,j)\in E} \left\{ \frac{1}{d_i^{out} d_j^{in}} \right\}. \tag{19}$$

Both the weakly and strongly directed forms of the von Neumann entropy (H_{VN}^{WD} and H_{VN}^{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 vertices linked by an edge. 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 (5) satisfies Passerini and Severini's conditions [13] for the density matrix. Moreover, we have shown that $\tilde{L}$ is Hermitian matrix, so its eigenvalues are all real. Hence theoretically, the density matrix interpretation of Passerini and Severini [13] can be extended to directed graphs. Second, when the out-degree and in-degree are the same at a vertex, then the von Neumann entropy for directed and undirected graphs are identical.

3 Heterogeneity Index and Commute Time

In this section, we present an index which quantifies the heterogeneous properties of directed graphs. We introduce the definitions of hitting time and commute time and describe how to compute them, then explore that on undirected graphs, there exists a relationship between heterogeneity index and commute time, and show that the similar relationship also applies to the directed graphs.

3.1 Heterogeneity Index of Directed Graphs

Following Estrada [10], in order to compute a heterogeneity index for directed graphs, we first require a local index to measure the irregularity associated with a single edge $(i,j) \in E$. Estrada [10] uses the following quantity to measure the variation in degree

$$\sigma_{ij}^{U} = [f(d_i) - f(d_j)]^2 \tag{20}$$

where $f(d)$ is a function of the vertex degree. To extend this measure to directed graphs, we measure the difference in out-degrees and in-degrees and write

$$\sigma_{ij}^{D} = [f(d_i^{out}) - f(d_j^{in})]^2. \tag{21}$$

This local heterogeneity measure takes on a value zero when the out-degree of the starting vertex is the same as the in-degree of the end vertex. On the other hand, the index should become larger when the difference of both degrees increases, thus we can select $f(d) = d^{-1/2}$. The local heterogeneity index associated with the irregularity of the edge $(i, j) \in E$ in a directed graph is given by

$$\sigma_{ij}^D = \left(\frac{1}{\sqrt{d_i^{out}}} - \frac{1}{\sqrt{d_j^{in}}} \right)^2. \tag{22}$$

To compute the global heterogeneity index of a directed graph we sum the local measure over all the edges in the graph to obtain

$$\rho^D(G) = \sum_{(i,j)\in E} \left\{ \frac{1}{\sqrt{d_i^{out}}} - \frac{1}{\sqrt{d_j^{in}}} \right\}^2 = \sum_{(i,j)\in E} \left\{ \frac{1}{d_i^{out}} + \frac{1}{d_j^{in}} \right\} - 2 \sum_{(i,j)\in E} \frac{1}{\sqrt{d_i^{out} d_j^{in}}}. \tag{23}$$

The heterogeneity index should take on a minimal value when the graph is regular, i.e. all the vertices have the same in-degree and out-degree. It is maximal when the graph is a star graph, i.e. there exists a central vertex such that all the other vertices connect and only connect to it. We calculate the lower and upper bounds of $\rho^D(G)$ according to these constraints. For a regular directed graph, suppose all the vertices have the same in-degree and out-degree d_0, then

$$\rho^D(G) = \sum_{(i,j)\in E} \left\{ \frac{1}{d_0} + \frac{1}{d_0} \right\} - 2 \sum_{(i,j)\in E} \frac{1}{d_0} = 0.$$

On the other hand, for a star graph, suppose that the central vertex has out-degree (in-degree) $|V| - 1$ and all the other vertices have in-degree (out-degree) 1. Then,

$$\rho^D(G) = \sum_{i=1}^{|V|} \left(\frac{1}{|V| - 1} + 1 \right) - 2 \sum_{i=1}^{|V|} \frac{1}{\sqrt{|V| - 1}} = \frac{|V|(|V| - 2\sqrt{|V| - 1})}{|V| - 1} \approx |V| - 2\sqrt{|V| - 1}.$$

We hence have the following lower and upper bounds for the heterogeneity index

$$0 \le \rho^D(G) = \sum_{(i,j)\in E} \left\{ \frac{1}{d_i^{out}} + \frac{1}{d_j^{in}} - \frac{2}{\sqrt{d_i^{out} d_j^{in}}} \right\} \le |V| - 2\sqrt{|V| - 1}. \tag{24}$$

Therefore we can define the normalized heterogeneity index of directed graphs as

$$\tilde{\rho}^D(G) = \frac{1}{|V| - 2\sqrt{|V| - 1}} \sum_{(i,j)\in E} \left\{ \frac{1}{d_i^{out}} + \frac{1}{d_j^{in}} - \frac{2}{\sqrt{d_i^{out} d_j^{in}}} \right\} \tag{25}$$

This index is zero for regular directed graphs, one for star graphs, i.e. $0 \le \tilde{\rho}^D(G) \le 1$.

3.2 Commute Time of Directed Graphs

To take our development one step further, we establish a relationship between the heterogeneity index and the commute time (or resistance distance) between nodes in a graph. To this end we commence by introducing the definitions of hitting time and commute time on directed graphs. The hitting time Q^D_{ij} is the expected number of steps for a random walk to reach vertex j for the first time, starting from vertex i. The commute time C^D_{ij} is the sum of Q^D_{ij} and Q^D_{ji}, i.e. $C^D_{ij} = Q^D_{ij} + Q^D_{ji}$, is the expected number of steps of a random walk starting at vertex i, visits j for the first time and then returns to vertex i.

Our expressions for both the hitting time and commute time are from Boley et al. [5]. We first introduce the definition of fundamental matrix Z which has elements

$$Z_{ij} = \sum_{t=0}^{\infty}(P^t_{ij} - \phi(j)), \quad 1 \leq i, j \leq |V| \tag{26}$$

or in matrix form,

$$Z = \sum_{t=0}^{\infty}(P^t - \mathbf{1}\phi) \tag{27}$$

where P is the transition matrix, $\mathbf{1} = (1, ..., 1)^T$ and ϕ is the stationary distribution.

The formulae for hitting time and commute time are

$$Q^D_{ij} = \frac{Z_{jj} - Z_{ij}}{\phi(j)}, \quad C^D_{ij} = Q^D_{ij} + Q^D_{ji} = \frac{Z_{jj} - Z_{ij}}{\phi(j)} + \frac{Z_{ii} - Z_{ji}}{\phi(i)}. \tag{28}$$

3.3 Relationship between Heterogeneity Index and Commute Time

According to Estrada [10], the normalized heterogeneity index of undirected graph has the following form

$$\tilde{\rho}^U(G) = \frac{1}{|V| - 2\sqrt{|V| - 1}} \sum_{(i,j)\in E} \left\{ \frac{1}{d_i} + \frac{1}{d_j} - \frac{2}{\sqrt{d_i d_j}} \right\}. \tag{29}$$

Recently, von Luxburg et al. [17] have shown that if the graph size is large enough, then the hitting time and commute time can be approximated by the resistance distance which takes on a simple form in terms of the vertex degree. In particular, $C^U_{ij} \approx vol\left(\frac{1}{d_i} + \frac{1}{d_j}\right)$ where vol is the volume of graph defined by $vol = \sum_{i=1}^{|V|} d_i$. As a result the first term appearing in the expression for Estrada's heterogeneity index can be expressed in terms of commute time.

To take this development one step further, we note that the normalized adjacency matrix for an undirected graph is given by $\tilde{A} = D^{-1/2}AD^{-1/2}$ where D is the diagonal matrix of vertex degrees. The normalized adjacency matrix has elements $\tilde{A}_{ij} = \frac{1}{\sqrt{d_i d_j}}$, if $(i, j) \in E$. As a result, in the heterogeneity index formula, if we make the substitutions $\frac{1}{d_i} + \frac{1}{d_j} = \frac{C_{ij}^U}{vol}$ and $\frac{1}{\sqrt{d_i d_j}} = \tilde{A}_{ij}$ we obtain the approximation

$$\tilde{\rho}^U(G) \approx \frac{1}{|V| - 2\sqrt{|V|-1}} \sum_{(i,j)\in E} \left\{ \frac{C_{ij}^U}{vol} - 2\tilde{A}_{ij} \right\}. \tag{30}$$

To extend this result to directed graphs, we note that

$$\sum_{(i,j)\in E} \left\{ \frac{1}{d_i^{out}} + \frac{1}{d_j^{in}} \right\} \approx \sum_{(i,j)\in E} \frac{C_{ij}^D}{vol} \tag{31}$$

where $vol = \sum_{i=1}^{|V|} d_i^{out} = \sum_{i=1}^{|V|} d_i^{in}$. If we denote by D_{out}, D_{in} the diagonal matrices of vertex out-degrees and in-degrees respectively, then the normalized adjacency matrix for a directed graph is $\tilde{A}^D = D_{out}^{-1/2} A D_{in}^{-1/2}$ with elements $\tilde{A^D}_{ij} = \frac{1}{\sqrt{d_i^{out} d_j^{in}}}$, if $(i,j) \in E$.

Hence, we obtain the following relationship between the heterogeneity index and commute time on directed graphs as

$$\tilde{\rho}^D(G) \approx \frac{1}{|V| - 2\sqrt{|V|-1}} \sum_{(i,j)\in E} \left\{ \frac{C_{ij}^D}{vol} - 2\tilde{A^D}_{ij} \right\}. \tag{32}$$

Thus we have shown that this relationship between heterogeneity index and commute time applies not only to undirected graphs but also to directed graphs.

Hence for both directed and undirected graphs, if the heterogeneity index is chosen in an appropriate way then there are two observations that can be drawn from this analysis. First, the heterogeneity index is proportional to the average commute time over pairs of nodes connected by an edge. Second, the heterogeneity index is greatest when the difference between the commute time and the twice the normalized adjacency matrix element is greatest. Hence, the heterogeneity index will be smallest for regular graphs and greatest for trees or star graphs.

4 Experiments and Evaluations

We have suggested several novel methods to measure the structural complexity of directed graphs. In this section, we aim to evaluate these methods on network data and give empirical analysis of their properties. First we examine both the weakly and strongly directed forms of von Neumann entropy, and compare their performance. Next, we explore whether our theoretically derived relationship between the heterogeneity index and commute time holds for both undirected and directed graphs.

4.1 The Datasets

Before undertaking our experiments, we first give a brief overview of the datasets used. The first dataset contains 150 undirected graphs in which the graph size varies from 50 to 100 nodes. Of this sample, 50 graphs are generated using the Erdos-Renyi model, which is considered as the most classical random graph model. An additional 50 graphs are generated according to the 'small-world' model, which was introduced by Watts and Strogatz [18]. The remaining 50 graphs are generated using the 'scale-free' model, which was developed by Barabasi and Albert [3]. The second dataset contains Protein-Protein Interaction (PPI) networks extracted from Franceschini et al. [11]. These graphs represent the interaction relationships between histidine kinase in different species of bacteria. The third dataset consists of 10 evolving directed networks. Each network commences from a fully connected network of size 5, and evolves gradually with new connections being established proportionally to the current dynamical activity of each vertex (preferential attachment). This dataset is generated using the model developed by Antiqueira et al. [2].

4.2 Entropy for Weakly and Strongly Directed Graphs

Equations (18) and (19) give the simplified forms of the von Neumann entropy for weakly and strongly directed graphs. We calculate them according to these two equations respectively and compare their behaviours with a reference entropy, i.e. the approximate von Neumann entropy generated using (15) (or equivalently, (16)), on the weakly and strongly directed networks in the third dataset.

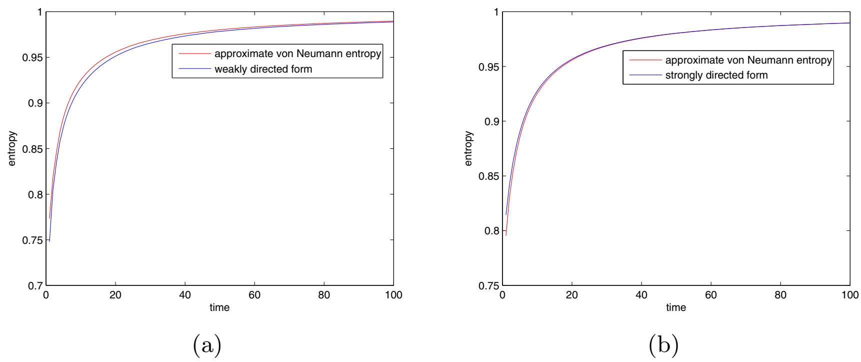


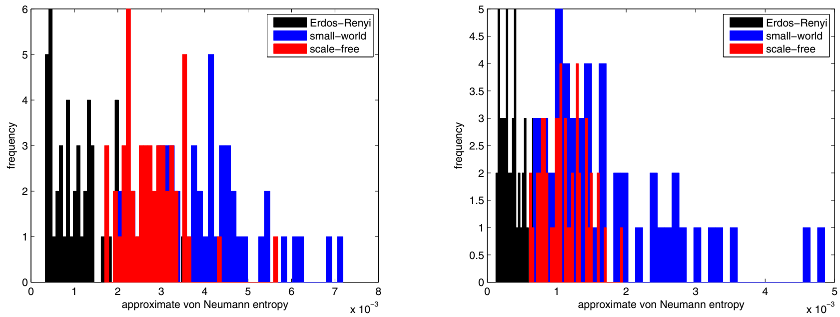
Fig. 1. Entropy for weakly & strongly directed graphs

We see in both Fig.1(a) and Fig.1(b), as the network evolves, both the simplified form and the reference entropy increase approximately monotonically until a plateaux value of unity is reached. Moreover, it is worth noting that the difference between these two quantities is negligible, thus we conclude that for weakly (strongly) directed graphs, the approximate von Neumann entropy and the simplified weakly (strongly) directed form we suggested are approximately equivalent.

We then explore whether the von Neumann entropy can be used to distinguish different types of graph. To this end we create directed versions of the Erdos-Renyi, 'small-world' and 'scale-free' graphs by deleting at random elements from the adjacency matrix. This has the effect of creating directed edges. In this analysis we consider the quantity

$$J = \left| H_{VN}^{D} - (1 - \frac{1}{|V|}) \right| = \frac{1}{2|V|^2} \left\{ \sum_{(i,j)\in E_2} \frac{1}{d_i^{out} d_j^{out}} + \sum_{(i,j)\in E} \frac{d_i^{in}}{d_j^{in} d_i^{out^2}} \right\}$$

which removes some of the size dependence of the entropy.



(a) (b)

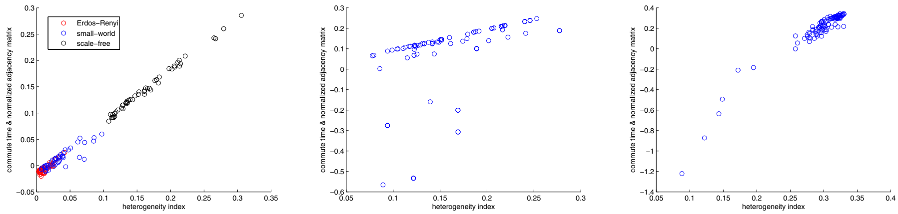
Fig. 2. Directed/Undirected graph characterization using von Neumann Entropy

In the left-hand of Fig.2, the plot shows superimposed histograms of J for each of the three types of directed graph. The main feature to note is that the Erdos-Renyi graphs are well separated from the 'small-world' and 'scale-free' graphs. Moreover, the 'scale-free' and 'small-world' networks although overlapped are reasonably well separated. The right-hand panel of Fig.2 repeats this analysis for the undirected versions of the three types of graph, using the original form of the von Neumann entropy suggested by Han et al. [12]. Here there is significantly more overlap, and the different types of network can not be easily separated, especially for the 'small-world' and 'scale-free' networks.

4.3 Heterogeneity Index and Commute Time

We have shown theoretically that the heterogeneity index has a linear dependance on the the commute time for both undirected and directed graphs. In this subsection we aim to confirm these results empirically. In Fig.3 we plot the heterogeneity index versus commute time for different types of graphs. Here the commute time of undirected graphs is calculated precisely using the graph spectral formula used by Qiu and Hancock [14]. Figure 3(a) shows the result for the Erdos-Renyi, 'small-world' and 'scale-free' graphs (shown in different colours). All three types of graphs follow a linear trend (i.e. they satisfy our theoretical prediction), but populate different parts of the 'heterogeneity-commute time' space. The second plot is for the protein-protein interaction networks. Although there are some outliers, most of the data falls in a linear regression curve. In fact, these outliers represent the graphs with particularly small graph size (e.g. 6 or 8), which is too small compared with others, thus these graphs do not perform the similar relation as other graphs do. Then we turn our attention to Fig.3(c), which is the plot of heterogeneity index versus average commute time for the directed graphs in the evolving sequence. The commute time here is computed according to (28). For the tightly clustered points in the upper right-hand corner of the plot, there is again a clear linear relationship, which confirms our theoretical prediction in (32).

Finally we explore the performance of directed graph characterization using the heterogeneity index. The histogram of the directed graph heterogeneity index is shown in Fig.4. In the histogram the 'scale-free' graphs are perfectly



(a) (b) (c)

Fig. 3. Relationship between Heterogeneity Index and Commute Time on undirected/directed graphs

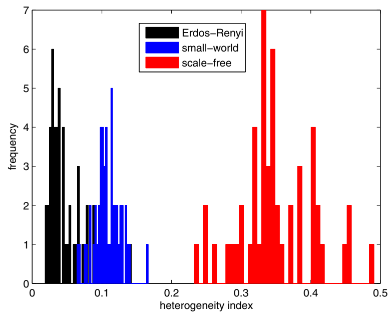


Fig. 4. Directed graph characterization using Heterogeneity Index

separated from the Erdos-Renyi and 'small-world' graphs. The result is not unexpected since for 'scale-free' graphs, the difference in the vertex in-degrees and out-degrees is particularly large, and the heterogeneity index of such graphs is greater than that for other types of graphs.

5 Conclusion

In this paper, motivated by the aim of developing novel and effective methods for quantifying the structural complexity of directed graphs, first we have developed approximations of the von Neumann entropy for both strongly and weakly directed graphs. They both depend on the vertex in-degree and out-degree statistics. Our approximations are based on using Chung's definition of normalized Laplacian matrix of directed graphs and simplifying the calculation via replacing the Shannon entropy by the quadratic entropy. Next, following the idea of developing the heterogeneity index for undirected graphs proposed by Estrada [10], we construct a similar measure which quantifies the heterogeneous properties of directed graphs. Moreover, concerning the commute time (or resistance distance), we have found that on an undirected graph, the heterogeneity index has a particular relation with it. Extending this correlation to directed graphs, we have discovered that they also exhibit a similar behaviour, which shows that the heterogeneity index can be approximated by the commute time and the normalized adjacency matrix. Then, in order to evaluate these methods and analyze their properties, we have undertaken some experiments on both undirected and directed network data and the experimental outcomes have demonstrated the effectiveness of our methods. In the future, our work can be extended by introducing more approaches to improving the measures we proposed in this paper for representing the structural complexity for directed graphs, and developing more novel indices which can reflect a directed graph's structure based on the entropy and heterogeneity index.

Acknowledgements. The authors are grateful to FAPESP (12/50986-7) and the University of York for financial support.

References

1. Amancio, D.R., Oliveira Jr., O.N., Costa, L.da F.: On the Concepts of Complex Networks to Quantify the Difficulty in Finding the Way Out of Labyrinths. Physica A 390, 4673–4683 (2011)
2. Antiqueira, L., Rodrigues, F.A., Costa, L.da F.: Modeling Connectivity in Terms of Network Activity. Journal of Statistical Mechanics: Theory and Experiment 0905.4706 (2009)
3. Barabasi, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509–512 (1999)
4. Berwanger, D., Gradel, E., Kaiser, L., Rabinovich, R.: Entanglement and the Complexity of Directed Graphs. Theoretical Computer Science 463, 2–25 (2012)

5. Boley, D., Ranjan, G., Zhang, Z.: Commute Times for a Directed Graph Using an Asymmetric Laplacian. Linear Algebra and Its Applications 435, 224–242 (2011)
6. Chung, F.: Laplacians and the Cheeger Inequailty for Directed Graphs. Annals of Combinatorics 9, 1–19 (2005)
7. Costa, L.da F., Rodrigues, F.A.: Seeking for Simplicity in Complex Networks. Europhysics Letters 85, 48001 (2009)
8. Escolano, F., Hancock, E.R., Lozano, M.A.: Heat Diffusion: Thermodynamic Depth Complexity of Networks. Physical Review E 85, 036206 (2012)
9. Escolano, F., Bonev, B., Hancock, E.R.: Heat Flow-Thermodynamic Depth Complexity in Directed Networks. In: Gimel'farb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR&SPR 2012. LNCS, vol. 7626, pp. 190–198. Springer, Heidelberg (2012)
10. Estrada, E.: Quantifying Network Heterogeneity. Physical Review E 82, 066102 (2010)
11. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., Jensen, L.J.: STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41, D808–D815 (2013)
12. Han, L., Escolano, F., Hancock, E.R., Wilson, R.C.: Graph Characterizations from Von Neumann Entropy. Pattern Recognition Letters 33, 1958–1967 (2012)
13. Passerini, F., Severini, S.: The Von Neumann Entropy of Networks. International Journal of Agent Technologies and Systems, 58–67 (2008)
14. Qiu, H., Hancock, E.R.: Clustering and Embedding Using Commute Times. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 1873–1890 (2007)
15. Ren, P., Wilson, R.C., Hancock, E.R.: Graph Characterization via Ihara Coefficients. IEEE Transactions on Neural Networks 22, 233–245 (2011)
16. Riis, S.: Graph Entropy, Network Coding and Guessing Games. The Computing Research Repository 0711.4175 (2007)
17. von Luxburg, U., Radl, A., Hein, M.: Hitting and Commute Times in Large Graphs are often Misleading. Data Structures and Algorithms 1003.1266 (2010)
18. Watts, D.J., Strogatz, S.H.: Collective Dynamics of 'Small-World' Networks. Nature 393, 440–442 (1998)
19. Xiao, B., Hancock, E.R., Wilson, R.C.: Graph Characteristics from the Heat Kernel Trace. Pattern Recognition 42, 2589–2606 (2009)