Investigating Network Classes by Measuring Their Complexity

Matthias Dehmer, Center for Mathematics, University of Coimbra, Apartado 3008, 3001-454 Coimbra, Portugal, dehmer@math.uc.pt

Stephan Borgert, Darmstadt University of Technology, Telecooperation Group, Hochschulstr. 10, D-64289 Darmstadt, Germany, borgert@tk.informatik.tu-darmstadt.de

Frank Emmert-Streib, Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, UK, v@bio-complexity.com

Abstract: In this paper, we propose an information-theoretic approach to discriminate graph classes structurally. For this, we use a measure for determining the structural information content of graphs. This complexity measure is based on a special information functional that quantifies certain structural information of a graph. To demonstrate that the entropy measure captures structural information meaningfully, we interpret some numerical results.

Key words: Network Modelling, Network Complexity Measures, Entropy

1 Introduction

Exploring quantitative measures for detecting network complexity has been a fascinating research topic since many decades. One important starting point was applying SHANNONS's information theory for investigating living systems, e.g., biological and chemical networks [9, 11, 3, 7, 14]. After this, methods to quantify the structural information content of graphs became quite popular where the classical ones were developed by, e.g., [12, 14, 10, 1]. In this paper, we deal with structurally discriminating network classes [5, 8] by using an information-theoretic technique for determining the structural information content of graphs. We define the structural information content of a given graph as its topological entropy. The resulting information measure will be interpreted as a graph complexity measure. Classical methods to measure the structural information content of graphs are often related to the problem to determine a partitioning of the underlying vertex set for obtaining a certain probability distribution [12, 14, 10]. For example, RASHEVSKY [12] defined the entropy of directed/undirected and unweighted graphs by partitioning the vertices in sets of indistinguishable vertices according to their dependence on local and non-local degree-dependencies. Then a probability distribution was obtained [12] by assigning a probability to each partition determined as the fraction of vertices

in this partition divided by the total number of vertices. Finally, this approach has been developed further in [10].

In this paper, we use a newly proposed entropy measure [4] to discriminate graph classes structurally. In contrast to the mentioned classical graph entropy methods, this measure is not based on determining vertex partitionings. The main construction principle is to assign a probability value to every vertex in a graph by using a certain information functional. Such a functional quantifies structural information of the graph under consideration. Based on numerical results, i.e., to calculate cumulative entropy distributions, we demonstrate that this entropy measure can capture important structural information of graphs by discriminating graph classes based on the structural information contents of involved graphs.

2 Information-Theoretic Complexity Measures for Graphs

In this section, we briefly outline the graph entropy method to measure the entropy of arbitrary undirected and connected networks [4]. Before starting, we express some mathematical preliminaries [2, 6, 4].

We start with the basic definition of a undirected, finite and connected graph G. We define G as $G = (V, E), |V| < \infty$, $E \subseteq \binom{V}{2}$. G is called connected if for arbitrary vertices v_i and v_j there exists an undirected path from v_i to v_j. Otherwise, we call G unconnected. In this paper, $\mathcal{G}_{UC}$ denotes the set of finite, undirected and connected graphs. The degree of a vertex $v \in V$ is denoted by $\delta(v)$ and equals the number of edges $e \in E$ which are incident with v. We call the quantity $\sigma(v) = \max_{u \in V} d(u, v)$ the eccentricity of $v \in V$, where $d(u, v)$ denotes the shortest distance between u and v. We want to notice that $d(u, v)$ is a metric. Further, $\rho(G) = \max_{v \in V} \sigma(v)$ is called the diameter of G. The j-sphere of a vertex v_i regarding $G \in \mathcal{G}_{UC}$ is defined as the set

$$S_j(v_i, G) := \{v \in V | \, d(v_i, v) = j, \, j \geq 1\}. \tag{1}$$

In order to define the entropy in general, let X be a discrete random variable with alphabet A and $p(x_i) = \Pr(X = x_i)$ be the probability mass function of X. Then, the entropy of X is defined as

$$H(X) := -\sum_{x_i \in A} p(x_i) \log(p(x_i)). \tag{2}$$

To repeat the novel graph entropy method recently introduced in [4], we first state the definition of a special information functional. Here, the information functional f^V quantifies structural information of a graph G by using metrical properties of graphs [13].

Definition 2.1 *Let $G \in \mathcal{G}_{UC}$ that is arbitrarily labeled. For a vertex $v_i \in V$, the information functional f^V is defined as*

$$f^V(v_i) := \alpha^{c_1|S_1(v_i,G)|+c_2|S_2(v_i,G)|+\cdots+c_{\rho(G)}|S_{\rho(G)}(v_i,G)|}, \quad c_k > 0, 1 \leq k \leq \rho(G), \alpha > 0. \tag{3}$$

c_k are real positive coefficients.

Definition 2.2 *The vertex probabilities are defined by the quantities*

$$p^V(v_i) := \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}. \tag{4}$$

Now, we define the structural information content of a graph $G \in \mathcal{G}_{UC}$ as its entropy of the underlying graph topology.

Definition 2.3 *Let $G = (V, E) \in \mathcal{G}_{UC}$. Then, we define the entropy of G by*

$$I_{f^V}(G) := -\sum_{i=1}^{|V|} p^V(v_i) \log(p^V(v_i)), \tag{5}$$

$$= -\sum_{i=1}^{|V|} \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)} \log\left(\frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}\right). \tag{6}$$

First, we want to remark that the process of defining information functionals and, hence the entropy of a graph by using structural properties or graph-theoretical quantities is not unique. We clearly see that each structural graph property or quantity captures certain structural information of an underlying graph differently. By considering Definition (2.1), we also observe that the information functional f^V contains the free parameter α and c_k. c_k can be used to weight structural characteristics of a graph in question. In terms of practical applications, the free parameter α can be determined, e.g., for graphs that should be classified, by applying an optimization method that optimizes α concerning known class labels of the graphs from an underlying training set. Then, the optimal α-value corresponds in this case to the parameter that leads us to the lowest classification error. Finally, we find that the value α can always be determined via an optimization procedure based on a given data set and, hence, is uniquely defined for a given classification problem [4].

3 Numerical Results

In this section, we demonstrate that our proposed entropy measure is suitable to discriminate graph classes structurally by using so called cumulative entropy distributions. For this, we here choose the class of so called ϑ-trees (VT) and a special class of rooted trees (RT). ϑ-trees are here defined as follows: $T_\vartheta = (V, E)$, $\vartheta \in \mathbb{N}$ is a rooted tree with the property that for the root $r \in V$ it holds $\delta(r) = \vartheta$. Further, for all internal vertices $v \in V$ it holds $\delta(v) = \vartheta + 1$ and leaves are vertices without successors. To obtain numerical results, we first define a special class of rooted trees: Each tree has the characteristic property that for $v \in V$ excluding leaves $1 \leq \delta(v) \leq 2$ holds. For generating the special class of ϑ-trees, we choose $\vartheta = 5$. Now, to determine the cumulative entropy distributions by varying the free parameters (α and c_k) we first express a definition.

Definition 3.1 *The graph classes C_α^{VT}, C_α^{RT} and the parameter sets $P_\mu(c_k)$ are generated in the following way:*

- *Starting from a fixed height h and ϑ, each $T_\vartheta \in C_\alpha^{VT}$ is created randomly by retaining the characteristic property of a ϑ-tree. Then, the graph entropy I_{f^V} is computed with the free parameter value α and a parameter setting $P_\mu(c_k)$.*

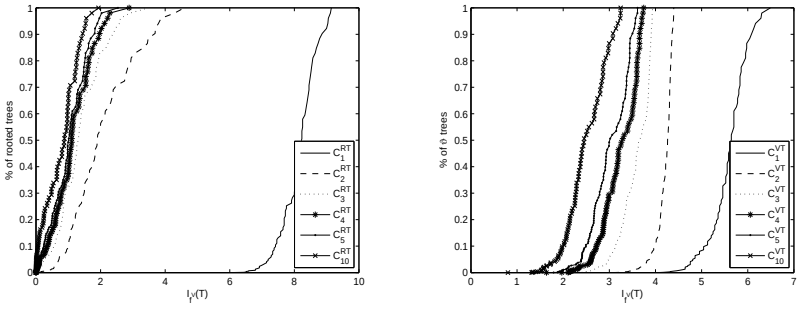

Figure 1: *Cumulative entropy distributions for* $C_1^{BT/VT}$-$C_5^{RT/VT}$ *and* $C_{10}^{RT/VT}$ *by using the parameter set* P_1.

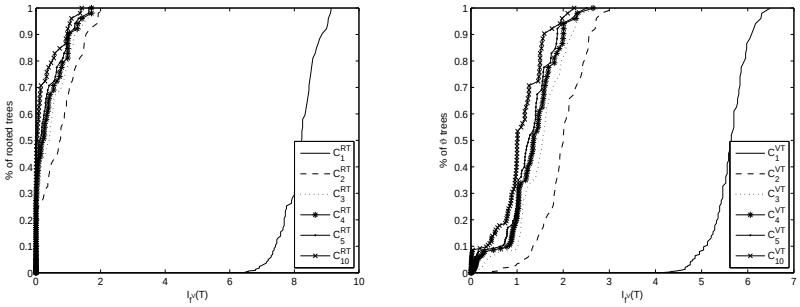

Figure 2: *Cumulative entropy distributions for* $C_1^{RT/VT}$-$C_5^{RT/VT}$ *and* $C_{10}^{RT/VT}$ *by using the parameter set* P_2.

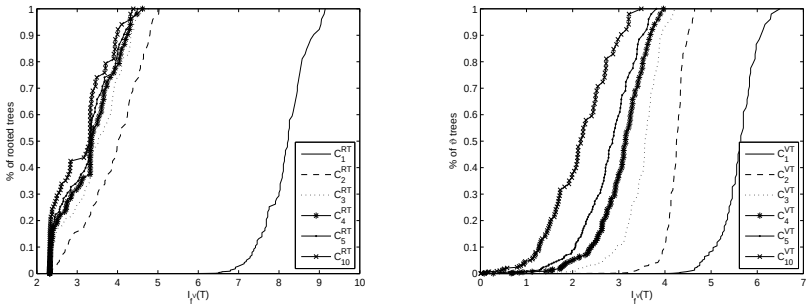

Figure 3: *Cumulative entropy distributions for* $C_1^{RT/VT}$-$C_5^{RT/VT}$ *and* $C_{10}^{RT/VT}$ *by using the parameter set* P_3.

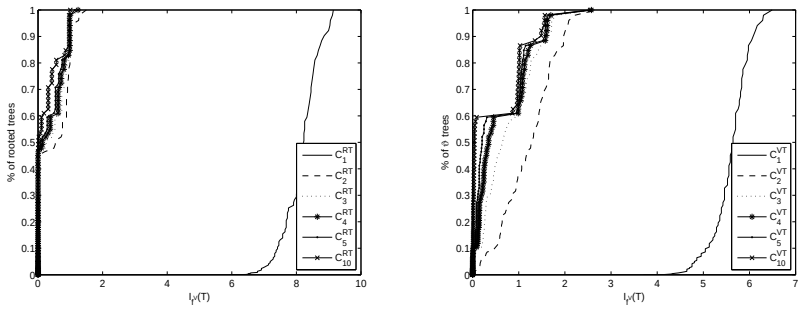

Figure 4: *Cumulative entropy distributions for* $C_1^{RT/VT}$-$C_5^{RT/VT}$ *and* $C_{10}^{RT/VT}$ *by using the parameter set* P_4.

- *Starting from a fixed height* h, *each* $T \in C_\alpha^{RT}$ *is created randomly by retaining the characteristic property as defined above. Then, the graph entropy* I_{f^V} *is computed with the free parameter value* α *and a parameter setting* $P_\mu(c_k)$.
- $P_1(c_k) = \{c_k \in \mathbb{R}_+ |\ c_1 := 2h,\ c_{k+1} := c_k - 1,\ k = 1, 2, \ldots, \rho(G) - 1\}$.
- $P_2(c_k) = \{c_k \in \mathbb{R}_+ |\ c_1 := 2h,\ c_2 := a \cdot c_1,\ c_3 := c_1 - 2,\ c_{k+1} := c_k - 1,\ a > 1,\ k = 3, 4, \ldots, \rho(G) - 3\}$.
- $P_3(c_k) = \{c_k \in \mathbb{R}_+ |\ c_1 := 1,\ c_2 := 2, \ldots, c_{\rho(G)} := \rho(G)\}$.
- $P_4(c_k) = \{c_k \in \mathbb{R}_+ |\ c_1 := 2h,\ c_2 := 2h - 1,\ c_3 := 2h - 2, c_k = 0,\ k > 3\}$.

We want to mention that using the parameter set P_1 means that we weight a stronger local branching in a graph. P_2 expresses that we weight paths of the 2-sphere more strongly than shorter ones. The meaning of P_3 is similarly to that of P_2. Finally, the definition of P_4 can be seen as an approximation of the information functional f^V because we set the cardinalities of the larger j-spheres equal to zero.

To interpret the cumulative entropy distributions for the generated tree classes (for $h = 8$) regarding their resulting entropies we look at Figure (1) - Figure (4). Here, the cumulative entropy distribution states the percentage rate of the total number of trees which possess an entropy value less or equal I_{f^V}. As a main result, we now find from Figure (1) - Figure (4) that for $\alpha \in \{1, 2, 3, 4, 5, 10\}$ the cumulative entropy distributions of C_α^{RT} are significantly different from the corresponding cumulative distributions of C_α^{VT}. We notice that the observation that the distribution for C_1^{RT} and C_1^{VT} seems to be almost equal is related to the fact that our entropy measure has always a maximum at $\alpha = 1$. This can be generally proven for an arbitrary undirected and connected graph. Putting it all together, the computed cumulative entropy distributions imply that in the shown cases the entropy measure proposed in Section (2) is able to detect that special rooted trees and ϑ-trees manifest structurally different graph classes.

References

[1] D. Bonchev. *Information Theoretic Indices for Characterization of Chemical Structures*. Research Studies Press, Chichester, 1983.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley & Sons, 2006.

[3] S. M. Dancoff and H. Quastler. Information content and error rate of living things. In H. Quastler, editor, *Essays on the Use of Information Theory in Biology*, pages 263–274. University of Illinois Press, 1953.

[4] M. Dehmer. A novel method for measuring the structural information content of networks. *Cybernetics and Systems*, 2008. in press.

[5] F. Emmert-Streib. The chronic fatigue syndrome: A comparative pathway analysis. *Journal of Computational Biology*, 14(7), 2007.

[6] F. Harary. *Graph Theory*. Addison Wesley Publishing Company, 1969.

[7] H. Linshitz. The information content of a battery cell. In H. Quastler, editor, *Essays on the Use of Information Theory in Biology*. University of Illinois Press, 1953.

[8] Alexander Mehler and Angelika Storrer. What are ontologies good for? evaluating terminological ontologies in the framework of text graph classification. In Uwe Mönnich and Kai-Uwe Kühnberger, editors, *OTT'06. Ontologies in Text Technology: Approaches to Extract Semantic Knowledge from Structured Information*, Publications of the Institute of Cognitive Science (PICS), pages 11–18, Osnabrück, 2007.

[9] H. Morowitz. Some order-disorder considerations in living systems. *Bull. Math. Biophys.*, 17:81–86, 1953.

[10] A. Mowshowitz. Entropy and the complexity of the graphs I: An index of the relative complexity of a graph. *Bull. Math. Biophys.*, 30:175–204, 1968.

[11] H. Quastler. *Information Theory in Biology*. University of Illinois Press, 1953.

[12] N. Rashevsky. Life, information theory, and topology. *Bull. Math. Biophys.*, 17:229–235, 1955.

[13] V. A. Skorobogatov and A. A. Dobrynin. Metrical analysis of graphs. *MATCH*, 23:105–155, 1988.

[14] E. Trucco. A note on the information content of graphs. *Bulletin of Mathematical Biology*, 18(2):129–135, 1956.