

The Usefulness of Topological Indices

Yuede Ma^{a,*}, Matthias Dehmer^{a,b,d,c,*}, Urs-Martin Künzi^b, Shailesh Tripathi^{h,e}, Modjtaba Ghorbani^g, Jin Tao^c, Frank Emmert-Streib^{e,f}

^a*School of Science, Xian Technological University, Xian, Shaanxi 710021, China*

^b*Swiss Distance University of Applied Science, Department of Computer Science, 3900 Brig, Switzerland*

^c*College of Artificial Intelligence, Nankai University, Tianjin 300350, China*

^d*Department of Biomedical Computer Science and Mechatronics, The Health and Life Science University-UMIT, A-6060 Hall in Tyrol, Austria*

^e*Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University, Tampere 33100, Finland*

^f*Institute of Biosciences and Medical Technology, Tampere 33520, Finland*

^g*Department of Mathematics, Faculty of Science, Shahid Rajaei, Teacher Training University, Tehran 16785-136, Iran*

^h*Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Steyr School of Management, Austria*

Abstract

A huge number of topological graph measures have been defined and investigated. It turned out that various graph measures failed to solve problems meaningfully in the context of characterizing graphs. Reasons for this range from selecting redundant and unfavorable graph invariants and the fact that many of those measures have been defined in an unreflected manner. In this paper, we extend the debate in the literature to find useful properties of structural graph measures. For this, we investigate the usefulness of topological indices for graphs quantitatively by assigning a feature vector to graph that contains 'useful' properties represented by certain measures. We show examples and compare the usefulness by using this apparatus based on distance measures and on an agglomerative clustering task.

Keywords: Quantitative Graph Theory, Networks, Topological Indices, Graphs, Topological Graph Measures, Data Science

2010 MSC: 62D99, 05C75, 68R10, 90B10

1. Introduction

Topological graph measures to characterize graphs quantitatively have been used for several decades. Structural graph measures can be divided into several categories, e.g., information-theoretic measures [23] and non-information-theoretic measures [26, 46]. These can be further subdivided by other categories which express what kind of graph invariant should be used; for instance, vertex degrees, distances in graphs, eigenvalues etc., see [46]. Another categorization of graph measures can be done by investigating their structural interpretation. As a result, symmetry measures [6, 16, 41], cyclicity measures [50] and branching measures [11, 15] have been investigated extensively. Therefore it's not surprising that a huge number of topological graph measures have been developed and applied in several scientific disciplines such as computer science [1], mathematical chemistry [17, 38], bioinformatics [12, 14] and so forth.

In fact, many structural graph measures have been not been investigated thoroughly. So, it appears that various graph measures are trivially related to each other, they possess a high

*contributed equally

Email addresses: mayuede0000@163.com (Yuede Ma), matthias.dehmer@umit.at (Matthias Dehmer), urs-martin.kuenzi@ffhs.ch (Urs-Martin Künzi), shailesh.tripathi@fh-steyr.at (Shailesh Tripathi), mghorbani@sru.ac.ir (Modjtaba Ghorbani), taoj@nankai.edu.cn (Jin Tao), frank.emmert-streib@tut.fi (Frank Emmert-Streib)

correlation score, and have very low discrimination power. Each of these mentioned aspects might be relevant in a particular discipline or to solve a special problem. For instance in [44, 9, 37] methods and features have been listed to define *useful* properties of graph measures. When thinking about the term *usefulness of topological graph measures*, several questions/problems can be raised, e.g.,

- Many features/properties such as discrimination power, correlation, structural sensitivity [31] etc. exist
- How to find/generate an exhaustive list of such features/properties?
- How should the term usefulness be defined at all?
- Can the usefulness be computed or measured?

In terms of originality, we mention to the best of our knowledge that we are the first to develop an approach to measure the usefulness of topological indices. The originality of our method lies in measuring the usefulness by using a comparative approach. Other and older contributions, e.g., [44, 9, 37] just discuss useful features; they don't measure anything.

The contribution of this paper is to come up with an quantitative approach to measure the usefulness of topological indices. We emphasize that this term/concept is not unique and is finally in the eye of a beholder. Our definition to quantify the usefulness of a topological graph measure with regard to a graph is based on a numerical feature vector, see Definition (2.5). This vector is composed by features/properties which contribute to the usefulness as they measure useful *effects* of those indices. In order to compare the usefulness of graph measures with regard to certain graph classes, we use distance measures and also define a clustering task and compare the optimal clustering solutions by using the quantity normalized mutual information (NMI) [35]. The results are summarized by the Tables (5,6).

The structure of the paper is as follows: In Section (1), we introduce the problem and discuss related work. Section (2) gives the definition of the usefulness and also important definitions of the used graph measures. Also, we give examples and use distance measures to compare the corresponding usefulness vectors, see Section (2). On top of that, we perform the clustering task, the statistical analysis and interpret the results in Section (2). The paper ends with a summary and conclusion in Section (3).

2. Methods and Results

Many topological graph measures also called topological indices have been defined, see, e.g., [6, 46, 26, 28]. A lot of these indices have been used for analyzing molecular graphs in structural chemistry, bioinformatics, drug design and related disciplines. From an application-oriented point of view, various topological indices turned out to be useful as they captured chemical or drug-related information significantly, see [5, 4, 3]. From a mathematical point of view, it's not always clear how the *usefulness* of a topological index should be defined, see [37]. In fact, various indices have been introduced in an unreflected manner. For instance, many indices are trivially related among each other; when applying those measures to data sets, it is expected that the output is highly correlated and does not give any additional insight. Also, it has only been little investigated whether topological graph indices possess useful properties. To shed light on this problem, quantities such as abruptness, smoothness and so forth have been introduced [31]. Furtula et al. [31] defined the so-called structure sensitivity and abruptness of degree-based indices and investigated properties thereof. Also, the so-called uniqueness of graph measures has been investigated extensively by Dehmer et al. [21, 20, 18]. This property relates to the problem to investigate how degenerate a measure is when discriminating graphs structurally [21, 36]. These properties mentioned above contribute to the usefulness of a topological index; such a list could be clearly extended. However, it's not straightforward to define a complete list of the mentioned properties that are useful to tackle this problem optimally.

By contrast, we here choose a quantitative approach for measuring the usefulness of an index. More precisely, we characterize the usefulness of a topological index I (structural graph measure) with regard to a graph G by a special feature vector. The basic idea is that the components of this vector describe *useful* properties. **We emphasize that the components of the feature vector cannot be learned by using Machine Learning-techniques.** Our definition uses features of graph structures based on our long standing experience in the field. As in many cases in structural graph analysis, the term usefulness cannot be defined uniquely and is in the eye of a beholder. Similar examples are the terms *structural complexity* [46] or *branching* [46] in graphs. Finally the mentioned components of the usefulness-vector (see Definition (2.5)) represent properties that are needed to describe useful features of topological graph measures. We explain one component of the usefulness-vector in Definition (2.5) namely $Abr(I, G)$ by way of example. $Abr(I, G)$ measures how abruptly the measured value I varies if applied to another graph (e.g., those whose graph structure are similar to each other). So, the usefulness of a graph measure I with high abruptness is surely questionable. Very similar arguments have been used to find the remaining components of the usefulness-vector in Definition (2.5). Anyway, to quantify and to define the usefulness remains a matter of definition and interpretation. This paper makes a first attempt to tackle this problem meaningfully.

We define our mathematical apparatus in the next section. In this section, we define important definitions to compare the usefulness of topological indices by using distance measures and by performing a clustering task.

2.1. Main Definitions

We start by defining graph measures required to define the usefulness, see also [31].

Definition 2.1. Let \mathcal{G} be a class of graphs and let $I : \mathcal{G} \rightarrow \mathbb{R}_+$ be a topological index. The Abruptness of I with regard to $G \in \mathcal{G}$ has been defined by [31]

$$Abr(I, G) = \max_{H \in \mathcal{G}} \left| \frac{I(H) - I(G)}{I(G)} \right|. \quad (1)$$

Definition 2.2. Let \mathcal{G} be a class of graphs and let $I : \mathcal{G} \rightarrow \mathbb{R}_+$ be a topological index. The structure sensitivity of I with regard to $G \in \mathcal{G}$ has been defined by [31]

$$SS(I, G) = \frac{1}{|\mathcal{G}|} \sum_{H \in \mathcal{G}} \left| \frac{I(H) - I(G)}{I(G)} \right|. \quad (2)$$

Definition 2.3. Let \mathcal{G} be a class of graphs and let $I : \mathcal{G} \rightarrow \mathbb{R}_+$ be a topological index. The uniqueness of an index I can be quantified by the so-called sensitivity measure regarding \mathcal{G} [36]:

$$S_{\mathcal{G}}(I) = \frac{|\mathcal{G}| - ndv}{|\mathcal{G}|}. \quad (3)$$

ndv is the number of graphs which can not be distinguished by I .

In case I is fully unique, we obtain $S_{\mathcal{G}}(I) = 1$, hence $ndv = 0$. The more degenerate the index I is, the smaller is $S(I)$. The uniqueness or degeneracy by employing various structural graph measures has been investigated extensively, see [24, 21, 19, 18, 36].

Definition 2.4. Let \mathcal{G} be a class of graphs and $G \in \mathcal{G}$. Let $D(G) = (d_{ij})_{ij}$ be the distance matrix of G ; d_{ij} is the shortest distance between the vertices i and j , see [33]. We define structural graph measure $I_{\lambda}(G)$ by

$$I_{\lambda}(G) = S_{\mathcal{G}}(I) \cdot \sum_{i=1}^k |\lambda_i^G|. \quad (4)$$

$\lambda_1^G, \lambda_2^G, \dots, \lambda_k^G$, $k \leq |V|$ are the non-zero eigenvalues of $D(G)$.

The idea of defining $I_\lambda(G)$ is to couple the uniqueness of I expressed by Equation (3) with a distinct measure namely $\sum_{i=1}^k |\lambda_i^G|$; the latter quantity represents the so-called *graph energy* [32, 39] of G by using the distance matrix.

Now we are able to define the usefulness of a topological index I with regard to a graph G . In order to motivate Definition (2.5), we emphasize that the quantities $I(G)$, $Abr(I, G)$, $SS(I, G)$, $I_\lambda(G)$ clearly contribute to the usefulness of an index; this is plausible because low abruptness ($Abr(I, G)$), good structure sensitivity ($SS(I, G)$), high uniqueness ($I_\lambda(G)$) are obviously desirable properties. Adding more vector components would quickly lead to redundancy. **A short discussion how to define the usefulness-vector can be found in the first part of Section (2).**

Definition 2.5. *Let \mathcal{G} be a class of graphs and let $I : \mathcal{G} \rightarrow \mathbb{R}_+$ be a topological index. Let $D(G)$ be the distance matrix of G . We define the usefulness of a topological index I with regard to a graph G by*

$$U^I(G) := (I(G), Abr(I, G), SS(I, G), I_\lambda(G)). \quad (5)$$

So, we now calculate the usefulness of any graph $G \in \mathcal{G}$. In this paper, we put the emphasis on undirected graphs.

2.2. Data

In order to perform our study we generate six data sets. These data sets are summarized by Table (1).

MS2265 contains 2265 selected chemical structures with different skeletons originating from the mass spectral database NIST [45]. This database has been already used in [25] for investigating different aspects of topological descriptors. It holds $4 \leq |V| \leq 19$; $2 \leq \text{diam}(G) \leq 15 \forall G \in \text{MS2265}$.

N_9 is the set of exhaustively generated pairwise non-isomorphic and connected graphs with 9 vertices [22]. T_{15} is the set of exhaustively generated pairwise non-isomorphic and connected trees with 15 vertices [22]. The tool Nauty [40] has been used to generate these two graph classes. Also, we generated scale-free networks by using the preferential attachment model [2] by varying $m = 1, \dots, 10$ and $p = \{1, \dots, 10\}$; m is the power of the preferential attachment and p is the number of edges to be added in each step. Towards the Erdős-Rényi networks, we randomly vary the total number of edges from 50 to 1200. Finally, we generate small-world networks due to Watts and Strogatz [48] by varying the neighborhood of the vertices of the lattice to be connected from 1 to 20, and use the rewiring probability $p_r = (0.0001, 0.99)$.

Data set	Description
MS2265	2265 small molecules
N_9	261080 pairwise non-isomorphic and connected graphs with 9 vertices.
T_{15}	7741 pairwise non-isomorphic and connected trees with 15 vertices.
Barabási game	150000 non-isomorphic scale-free networks with 50 vertices.
Erdős-Rényi	150000 non-isomorphic random networks with 50 vertices.
Small-world	150000 non-isomorphic networks with 50 vertices.

Table 1: Different graph classes used for the analysis.

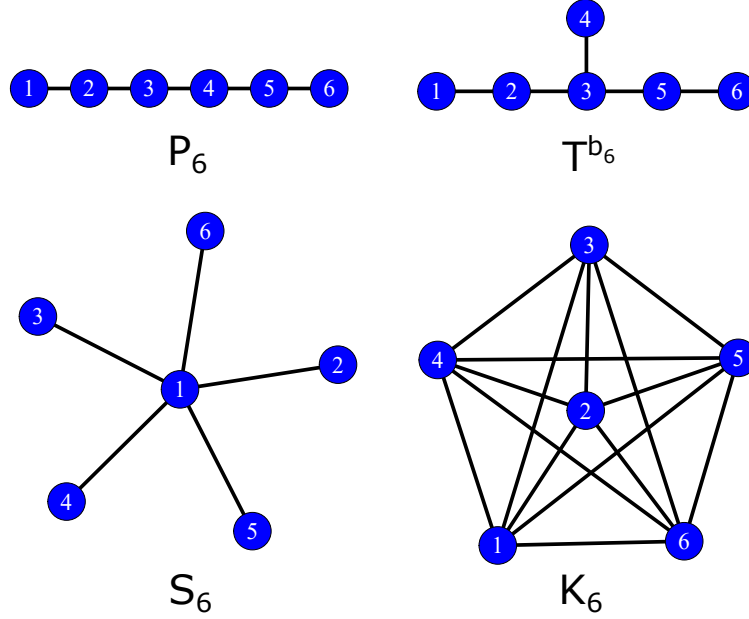


Figure 1: Four example graphs.

Short Form	Topological Index
info1	Information-Theoretic Measure (vertex degree) [13, 23, 42]
info2	Information-Theoretic Measure (sphere)[13, 23, 42]
info3	Information-Theoretic Measure (pathlength) [13, 23, 42]
info4	Information-Theoretic Measure (vertex centrality)[13, 23, 42]
wiener	Wiener index [49, 23, 42]
randić	Randić index [43, 23, 42]
topoinf	Topological Information Content [41, 23]

Table 2: Descriptors analyzed for sensitivity.

Graphs ↓	Wiener (W)				Randić (R)			
	$W(G)$	$Abr(W, G)$	$SS(W, G)$	$W_\lambda(G)$	$R(G)$	$Abr(R, G)$	$SS(R, G)$	$R_\lambda(G)$
P_6	35.000	0.571	0.351	0.865	2.914	0.233	0.029	16.002
T^{b_6}	31.000	0.516	0.270	0.767	2.808	0.204	0.039	14.195
S_6	25.000	0.400	0.141	0.613	2.236	0.342	0.282	11.341
K_6	15.000	1.333	0.514	0.357	3.000	0.255	0.044	6.607

Table 3: Usefulness vectors for the Wiener and Randić index with regard to the four example graphs.

2.3. Examples of the Usefulness and Comparison

Here we calculate the usefulness vectors for some example graphs, see Figure (1). To calculate these vectors, we use the Wiener and Randić index and compose the vectors based on Definition (2.5) for the shown graphs. The results are shown in Table (3). The vector components are the measures that contribute to the usefulness, see Definition (2.5). As already mentioned repeatedly, here the usefulness is not a single numerical value that can be computed by assuming a topological index I . Based on our vector approach, the usefulness becomes *measurable* by comparing these vectors. For this, many vector-based distance measures are basically applicable. In this paper, we employed the well-known *euclidian distance* [29] and we calculated a distance matrix, see Table (4).

	Graphs	Wiener (W)				Randić (R)			
		P_6	T^{b_6}	S_6	K_6	P_6	T^{b_6}	S_6	K_6
Wiener	P_6	0.000	4.002	10.007	20.022	35.480	34.846	34.399	32.514
	T^{b_6}	4.002	0.000	6.005	16.028	31.954	31.229	30.646	28.605
	S_6	10.007	6.005	0.000	10.054	26.919	26.019	25.166	22.803
	K_6	20.022	16.028	10.054	0.000	19.806	18.484	16.870	13.581
Randić	P_6	35.480	31.954	26.919	19.806	0.000	1.810	4.717	9.395
	T^{b_6}	34.846	31.229	26.019	18.484	1.810	0.000	2.924	7.591
	S_6	34.399	30.646	25.166	16.870	4.717	2.924	0.000	4.802
	K_6	32.514	28.605	22.803	13.581	9.395	7.591	4.802	0.000

Table 4: Euclidian distance between the usefulness vectors by using the Wiener and Randić index.

Let's consider the following example by choosing the Wiener index as topological index. Also, let's consider the two graphs P_6 and T^{b_6} . By using Equation (5), the computation of the usefulness vectors yields to

$$U^W(P_6) = (W(P_6), \text{Abr}(W, P_6), \text{SS}(W, P_6), I_\lambda(P_6)), \quad (6)$$

$$U^W(T^{b_6}) = (W(T^{b_6}), \text{Abr}(W, T^{b_6}), \text{SS}(W, T^{b_6}), I_\lambda(T^{b_6})). \quad (7)$$

For instance, we obtain $\text{Euclid}(U^W(P_6), U^W(T^{b_6})) = 4.002$, see Table(4). We also see that $\text{Euclid}(U^W(P_6), U^W(T^{b_6}))$ gives the lowest distance for W and, hence, the corresponding vectors are most similar. In general, if we assume that a distance measure vanishes, we then achieve maximum similarity, see [27]. In case we examine the values in Table (4) regarding the mixed vectors by taking W and R into account, we see the distance values are rather high (low similarity). So, the comparison of the usefulness vectors give the lowest values on the four example graphs by using the same topological index.

2.4. Comparative Analysis by Means of Clustering

In this section, we describe another comparative analysis of the usefulness of topological graph measures by means of clustering. The used topological graph indices can be seen in Table 2. Let $G = (V, E)$ be a graph; the information-theoretic measures [13, 23, 42] we use in this paper can be derived from the family

$$I_f(G) := - \sum_{i=1}^{|V|} \frac{f(v_i)}{\sum_{j=1}^n f(v_j)} \log \left(\frac{f(v_i)}{\sum_{j=1}^n f(v_j)} \right), \quad v_i \in V, \quad (8)$$

by utilizing special information functionals $f : V \rightarrow \mathbb{R}_+$ as indicated in Table (2). As defined in Section (2.1), the usefulness of a graph measure with regard to a graph is defined by a vector that contains features which contribute to quantify the usefulness, see Equation (5).

By using different graph classes, we perform a clustering task and, finally, we compare the clustering solutions by utilizing the well-known measure *normalized mutual information* (NMI) [47]. So, two structural graph measures have similar uniqueness if the clustering solutions are similar measured by NMI [47]. The similarity of the clustering solutions will be here measured by computing a p and q -value to determine the significance.

In order to start describing the clustering analysis, we note that all values of the calculated topological indices are normalized concerning the interval $[0, 1]$. Because some of the graph classes discussed in Section (2.2) are rather large, the computational complexity is expected to be high. We mention that we used a large computer cluster to perform the analysis described below and the computations took several days. Therefore it makes sense to calculate random samples to overcome this problem. So, we calculate $s_k = 500, 1000, 2000$ samples of the graph classes in Table (1) and apply *agglomerative clustering* [35]. To perform the clustering, we applied Euclidean distance [27] to the obtained vectors by using Definition (2.5). For calculation the vectors on

the used graphs (see Table (1)), we take the measures into account which are shown by Table (2). When using agglomerative clustering, we employ the *silhouette coefficient* [35] to optimize the number of clusters denoted by $\text{CL}(s_k, I) \in \{2, 3, \dots, 100\}$; $I : \mathcal{G} \rightarrow \mathbb{R}_+$ is a topological index. Finally, we compare two clustering solutions resulting from using the indices I_i and I_j for each number s_k by utilizing normalized mutual information:

$$\text{NMI}_{I_i, I_j}^{\mathcal{G}} = \text{NMI}^{\mathcal{G}}(\text{CL}(s_k, I_i), \text{CL}(s_k, I_j)). \quad (9)$$

These clustering solutions are obtained by using the silhouette coefficient to obtain the best partitions; we repeat the analysis 100 times for each graph class \mathcal{G} (see Table 1) and compute the average of NMI.

We apply a statistical test to determine the significance of $\text{NMI}_{I_i, I_j}^{\mathcal{G}}$ by defining the null hypothesis:

$$H_0 : \text{NMI}_{I_i, I_j}^{\mathcal{G}} = \text{NMI}_{I_i, I_j}^{\mathcal{G}, \text{random}} \quad (10)$$

$$H_1 : \text{NMI}_{I_i, I_j}^{\mathcal{G}} > \text{NMI}_{I_i, I_j}^{\mathcal{G}, \text{random}} \quad (11)$$

We define $C(I_i) := C_1, C_2, \dots, C_{k_1}$, and $C(I_j) = C_1, C_2, \dots, C_{k_2}$. These are partitions we obtain by applying the clustering for I_i and I_j for n graphs. To perform the hypothesis testing procedure, we apply the following steps:

1. Calculate $\text{NMI}_{I_i, I_j}^{\mathcal{G}} = \text{NMI}(C(I_i), C(I_j))$ between two clustering solutions $C(I_j)$ and $C(I_i)$.
2. Randomize the partitions $C(I_i)$, $C(I_j)$ to obtain a randomized partition $C^{\text{random}}(I_i) = \text{rand}(C(I_i))$ and $C^{\text{random}}(I_j) = \text{rand}(C(I_j))$.
3. Compute $\text{NMI}_{I_i, I_j}^{\mathcal{G}, \text{random}} = \text{NMI}(C^{\text{random}}(I_i), C^{\text{random}}(I_j))$.
4. Repeat Step 2 and 3 10000 times.
5. Calculate the p -value,

$$p = \frac{\# \left(\text{NMI}_{I_i, I_j}^{\mathcal{G}} < \{ \text{NMI}_{I_i, I_j}^{\mathcal{G}, \text{random}} \}_{i=1}^{10000} \right)}{10000}. \quad (12)$$

Also, we perform the multiple testing correction due to Bonferroni [34] to obtain adjusted p -values. Here, we use the threshold $\alpha = 0.01$ on the adjusted p -values to select the significant normalized mutual information values of pairs of the used topological indices.

The numerical results of the clustering analysis are shown by the Tables (5), (6). We start interpreting the results by Table (5) that shows the results from the clustering comparison for MS2265 and for the two classes of exhaustively generated networks N_9 and T_{15} . Regarding MS2265 the index pair gives the highest q -value. The second highest q -value also involves the computation of the measure info4. We also see that info1 fails to produce significant q -values. The results when comparing the number of NMI for which the q -values are significant are quite similar for the three graph classes shown by Tables (5). In this paper, we defined the usefulness of a topological graph measures with regard to a graph by a numerical feature vector, see Equation (5). These features capture 'useful' properties of a graph measure. If we obtain a significant q -value from the clustering comparison by using I_i and I_j , this means that the usefulness of I_i and I_j are similar based on the clustering ability of these indices. By contrast, the usefulness of two indices is dissimilar if the NMI value is very little and the q -value is not significant. In Table (5), we see that several pairs of topological indices have quite different clustering ability based in case a graph is represented by the defined feature vector and, hence, their NMI-value is not similar.

The results by using classes of random graphs are shown in Table (6). We see that almost all pairs of the used topological indices produced significant q -values. A plausible reason for this result could be the larger size of the graphs and the fact that we only performed the clustering comparison on relatively small samples by keeping in mind that the population size of the random graph classes is huge. Also, the discrimination power of the computed measures in Equation (5)

	Pair of descriptors	MS2265			N_9			T_{15}		
		NMI	p -value	q -value	NMI	p -value	q -value	NMI	p -value	q -value
1	info2_info1	0.0552	0.0203	0.0223	0.0092	0.0073	0.0109	0.0570	0.0084	0.0104
2	info3_info1	0.0567	0.0205	0.0223	0.0113	0.0069	0.0109	0.0570	0.0084	0.0104
3	info4_info1	0.0575	0.0168	0.0223	0.0113	0.0123	0.0162	0.1413	0.0011	0.0028
4	wiener_info1	0.0456	0.0182	0.0223	0.0725	0.0014	0.0038	0.0805	0.0043	0.0075
5	randić_info1	0.0429	0.0166	0.0223	0.0210	0.0079	0.0111	0.0489	0.0084	0.0104
6	topoinf_info1	0.0281	0.0229	0.0229	0.0206	0.00001	0.0005	0.0930	0.0042	0.0075
7	info3_info2	0.8887	<0.00001	0.0002	0.0680	0.0012	0.0035	1.0000	<0.00001	0.0003
8	info4_info2	0.6334	<0.0000	0.0001	0.1796	0.0001	0.0009	0.1476	0.0012	0.0028
9	wiener_info2	0.5354	0.0086	0.0150	0.0425	0.0029	0.0061	0.0211	0.0204	0.0226
10	randić_info2	0.4324	0.0001	0.0005	0.1168	0.0007	0.0025	0.0038	0.0564	0.0564
11	topoinf_info2	0.1017	0.0212	0.0223	0.0057	0.0995	0.1161	0.0691	0.0080	0.0104
12	info4_info3	0.6533	<0.00001	0.0001	0.0991	0.0002	0.0011	0.1476	0.0012	0.0028
13	wiener_info3	0.5086	0.0085	0.0150	0.0287	0.0032	0.0062	0.0211	0.0204	0.0226
14	randić_info3	0.4411	0.0001	0.0005	0.0209	0.0059	0.0102	0.0038	0.0564	0.0564
15	topoinf_info3	0.1001	0.0210	0.0223	0.0045	0.1701	0.1880	0.0691	0.0080	0.0104
16	wiener_info4	0.4912	0.0006	0.0016	0.0854	0.0004	0.0016	0.1921	0.0006	0.0027
17	randić_info4	0.5658	0.0002	0.0007	0.1720	0.0001	0.0009	0.1162	0.0017	0.0035
18	topoinf_info4	0.1165	0.0055	0.0116	0.0048	0.1895	0.1989	0.1489	0.0010	0.0028
19	randić_wiener	0.4690	0.0002	0.0007	0.0733	0.0021	0.0049	0.3709	0.0002	0.0017
20	topoinf_wiener	0.1142	0.0125	0.0203	0.0036	0.2120	0.2120	0.2762	0.0003	0.0017
21	topoinf_randić	0.1235	0.0036	0.0084	0.0055	0.0209	0.0259	0.3920	0.0003	0.0017

Table 5: Comparison of different pair of topological indices. The significant pairs are highlighted in bold for which the q -values are less than threshold $\alpha < 0.01$.

	Pair of descriptors	Barabási			Erdős-Rényi			Small-world		
		NMI	p -value	q -value	NMI	p -value	q -value	NMI	p -value	q -value
1	info2_info1	0.2010	0.0001	0.0001	0.0065	0.0566	0.0661	0.3397	0.0002	0.0008
2	info3_info1	0.2890	0.0001	0.0001	0.2569	0.0001	0.0001	0.0040	0.0676	0.0710
3	info4_info1	0.0908	0.0001	0.0001	0.0064	0.0484	0.0598	0.0542	0.0013	0.0026
4	wiener_info1	0.1018	0.0001	0.0001	0.0526	0.0005	0.0008	0.2609	0.0003	0.0009
5	randić_info1	0.0087	0.0020	0.0025	0.0071	0.0397	0.0522	0.0295	0.0030	0.0042
6	topoinf_info1	0.1327	0.0001	0.0002	0.3932	0.0001	0.0001	0.6373	0.0001	0.0007
7	info3_info2	0.2411	0.0001	0.0001	0.0147	0.0002	0.0003	0.0067	0.0290	0.0320
8	info4_info2	0.0182	0.0001	0.0001	0.3708	0.0001	0.0001	0.1251	0.0011	0.0026
9	wiener_info2	0.0118	0.0001	0.0002	0.1977	0.0001	0.0001	0.4155	0.0001	0.0005
10	randić_info2	0.0104	0.0004	0.0005	0.2850	0.0001	0.0001	0.0657	0.0028	0.0041
11	topoinf_info2	0.1118	0.0001	0.0001	0.0032	0.2167	0.2167	0.1987	0.0014	0.0026
12	info4_info3	0.0626	0.0001	0.0001	0.0152	0.0001	0.0002	0.0115	0.0058	0.0076
13	wiener_info3	0.0623	0.0001	0.0001	0.1700	0.0001	0.0001	0.0432	0.0002	0.0008
14	randić_info3	0.0115	0.0001	0.0001	0.0183	0.0002	0.0003	0.0087	0.0184	0.0216
15	topoinf_info3	0.0788	0.0001	0.0001	0.1218	0.0001	0.0001	0.0028	0.1088	0.1088
16	wiener_info4	0.5723	0.0001	0.0001	0.2383	0.0001	0.0001	0.0902	0.0008	0.0024
17	randić_info4	0.0444	0.0004	0.0005	0.3250	0.0001	0.0002	0.2947	0.0001	0.0005
18	topoinf_info4	0.0194	0.0074	0.0078	0.0032	0.1986	0.2086	0.0325	0.0017	0.0030
19	randić_wiener	0.0024	0.2872	0.2872	0.1735	0.0001	0.0002	0.0504	0.0021	0.0033
20	topoinf_wiener	0.0248	0.0043	0.0050	0.0203	0.0068	0.0096	0.1829	0.0012	0.0026
21	topoinf_randić	0.0288	0.0047	0.0052	0.0038	0.1304	0.1441	0.0106	0.0185	0.0216

Table 6: Comparison of different pair of topological indices by sampling random networks for which $|V| = 50$. The significant pairs are highlighted in bold for which the q -values are less than threshold $\alpha < 0.01$. NMI is calculated of clustering solutions by using 2 indices by optimizing the best partition using the silhouette measure.

is probably higher on small samples of random graphs compared to MS2265 and the exhaustively generated graph classes. So, it seems to be more challenging to find pairs of topological indices with significant q -values on smaller graphs where the graphs are structurally similar. Generating and interpreting such results would involve using graph similarity or distance measures [30].

3. Summary and Conclusion

In this paper, we defined the usefulness of topological indices with regard to a graph by a numerical feature vector. In earlier paper, several attempts were made to define properties of *good* topological graph measures, e.g., see [8, 9, 10, 7]. For instance, features properties were listed that contribute to quantify topological complexity meaningfully. The question of whether a structural graph measure is useful or not is somewhat related to the latter problem. In this paper, we chose a quantitative rather than a qualitative approach. In Section (2.3), we calculated the usefulness on four example graphs and compared those vectors by using the euclidian distance. It would be very

interesting to find more topological indices resulting in low distance values by using appropriate graph classes. Then, we would have found graphs and indices whose usefulness is highly similar based on the comparative analysis. A second step should involve finding other measures which are highly correlated with the (euclidian) distance. This could serve as another kind of interpretation of the usefulness.

We also compared the usefulness of pairs of topological indices by comparing clustering solutions by using these indices and normalized mutual information [35]. Some of the used descriptors performed well, i.e., their q -values of the clustering comparisons were significant and, hence, the uniqueness of the involved measures can be regarded as similar.

Future work involves seeking pairs of indices that produce significant q -values by using our clustering apparatus. Also, other graph classes could be used and the NMI-values can be compared. Also, we like to explore other definitions and methods to defined and compare the usefulness of topological graph measures. A possible research direction for this problem area would involve using machine learning techniques by solving optimization problem. In such a case, the usefulness could be defined based solving optimization problem, e.g., learning parameters of the involved methods optimally.

4. Acknowledgments

Matthias Dehmer thanks the Austrian Science Funds for supporting this work (project P30031).

References

- [1] E. B. Allen. Measuring graph abstractions of software: An information-theory approach. In *Proceedings of the 8-th International Symposium on Software Metrics table of contents*, page 182. IEEE Computer Society, 2002.
- [2] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [3] S. C. Basak and V. R. Magnuson. Molecular topology and narcosis. *Arzeim.-Forsch./Drug Design*, 33(I):501–503, 1983.
- [4] S. C. Basak, S. Majumdar, G. Grundwald, and A. Nandy. Mathematical descriptors of molecules and biomolecules: Development and applications to characterization of chemical libraries, qsar, drug design, nanotoxicology, and zika peptide vaccine design. In *Proceedings of the Conference BIO NANO MATH CHEM: NANO SCIENCE IN CHEMISTRY, PHYSICS. BIOLOGY AND MATHEMATICS*, Cluj, Romania, 2017.
- [5] S. C. Basak, Sonja Nikolić, Nenad Trinajstić, Dragan Amic, and Drago Beslo. QSPR modeling: Graph connectivity indices versus line graph connectivity indices. *J. Chem. Inf. Comput. Sci.*, 40(4):927–933, 2000.
- [6] D. Bonchev. *Information Theoretic Indices for Characterization of Chemical Structures*. Research Studies Press, Chichester, 1983.
- [7] D. Bonchev. *Complexity in Chemistry. Introduction and Fundamentals*. Taylor and Francis, 2003. Boca Raton, FL, USA.
- [8] D. Bonchev, O. Mekenyan, and N. Trinajstić. Isomer discrimination by topological information approach. *J. Comp. Chem.*, 2(2):127–148, 1981.
- [9] D. Bonchev and O. E. Polansky. On the topological complexity of chemical systems. In R. B. King and D. H. Rouvray, editors, *Graph Theory and Topology*, pages 125–158. Elsevier, 1987. Amsterdam, The Netherlands.

- [10] D. Bonchev and D. H. Rouvray. *Complexity in Chemistry, Biology, and Ecology*. Mathematical and Computational Chemistry. Springer, 2005. New York, NY, USA.
- [11] D. Bonchev. Topological order in molecules 1. Molecular branching revisited. *Journal of Molecular Structure: THEOCHEM*, 336(2-3):137 – 156, 1995.
- [12] J. C. Claussen. Offdiagonal complexity: A computationally quick network complexity measure - Application to protein networks and cell division. In A. Deutsch, R. Bravo de la Parra, R. de Boer, O. Diekmann, P. Jagers, E. Kisdi, M. Kretzschmar, P. Lansky, and H. Metz, editors, *Mathematical Modeling of Biological Systems, Volume II*, pages 303–311. Birkhäuser, Boston, 2007.
- [13] M. Dehmer. Information processing in complex networks: Graph entropy and information functionals. *Appl. Math. Comput.*, 201:82–94, 2008.
- [14] M. Dehmer, N. Barbarini, K. Varmuza, and A. Graber. Novel topological descriptors for analyzing biological networks. *BMC Structural Biology*, 10(18), 2010.
- [15] M. Dehmer, Z. Chen, F. Emmert-Streib, A. Mowshowitz, Y. Shi, S. Tripathi, and Y. Zhang. Towards detecting structural branching and cyclicity in graphs: A polynomial-based approach. *Information Sciences*, 471:19–28, 2019.
- [16] M. Dehmer, Z. Chen, F. Emmert-Streib, A. Mowshowitz, K. Varmuza, L. Feng, H. Jodlbauer, Y. Shi, and J. Tao. The orbit-polynomial: A novel measure of symmetry in networks. *IEEE Access*, 8:36100–36112, 2020.
- [17] M. Dehmer and F. Emmert-Streib. Structural information content of networks: Graph entropy based on local vertex functionals. *Computational Biology and Chemistry*, 32:131–138, 2008.
- [18] M. Dehmer, F. Emmert-Streib, and M. Grabner. A computational approach to construct a multivariate complete graph invariant. *Information Sciences*, 260:200–208, 2014.
- [19] M. Dehmer and M. Grabner. The discrimination power of molecular identification numbers revisited. *MATCH Commun. Math. Comput. Chem.*, 69(3):785–794, 2013.
- [20] M. Dehmer, M. Grabner, A. Mowshowitz, and F. Emmert-Streib. An efficient heuristic approach to detecting graph isomorphism based on combinations of highly discriminating invariants. *Advances in Computational Mathematics*, 39:311–325, 2012.
- [21] M. Dehmer, M. Grabner, and K. Varmuza. Information indices with high discriminative power for graphs. *PLoS ONE*, 7:e31214, 2012.
- [22] M. Dehmer, M. Moosbrugger, and Y. Shi. Encoding structural information uniquely with polynomial-based descriptors by employing the randić matrix. *Applied Mathematics and Computation*, 268:164–168, 2015.
- [23] M. Dehmer and A. Mowshowitz. A history of graph entropy measures. *Information Sciences*, 1:57–78, 2011.
- [24] M. Dehmer, L. Müller, and A. Graber. New polynomial-based molecular descriptors with low degeneracy. *PLoS ONE*, 5(7), 2010.
- [25] M. Dehmer, K. Varmuza, S. Borgert, and F. Emmert-Streib. On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. *J. Chem. Inf. Model.*, 49:1655–1663, 2009.
- [26] J. Devillers and A. T. Balaban. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers, 1999. Amsterdam, The Netherlands.

- [27] M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer; 2nd ed., 2012.
- [28] M. V. Diudea, I. Gutman, and L. Jäntschi. *Molecular Topology*. Nova Publishing, 2001. New York, NY, USA.
- [29] F. Emmert-Streib, M. Dehmer, and S. Moutari. *Mathematical Foundations of Data Science Using R*. De Gruyter Oldenbourg, 2020.
- [30] F. Emmert-Streib, M. Dehmer, and Y. Shi. Fifty years of graph matching, network alignment and network comparison. *Information Sciences*, 346-347:180–197, 2016.
- [31] B. Furtula, I. Gutman, and M. Dehmer. On structure-sensitivity of degree-based topological indices. *Applied Mathematics and Computation*, 219:8973–8978, 2013.
- [32] I. Gutman. The energy of a graph: Old and new results. In A. Betten, A. Kohnert, R. Laue, and A. Wassermann, editors, *Algebraic Combinatorics and Applications*, pages 196–211. Springer Verlag, 2001. Berlin.
- [33] F. Harary. *Graph Theory*. Addison Wesley Publishing Company, 1969. Reading, MA, USA.
- [34] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.
- [35] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.
- [36] E. V. Konstantinova. The discrimination ability of some topological and information distance indices for graphs of unbranched hexagonal systems. *J. Chem. Inf. Comput. Sci.*, 36:54–57, 1996.
- [37] V. Kraus, M. Dehmer, and F. Emmert-Streib. Probabilistic inequalities for evaluating structural network measures. *Information Sciences*, 228:220–245, 2014.
- [38] X. Li, Y. Shi, and I. Gutman. *Graph Energy*. Springer, 2012.
- [39] X. Li, Y. Shi, and I. Gutman. *Graph Energy*. Springer, 2012. New York.
- [40] B. D. McKay. Isomorph-free exhaustive generation. *Journal of Algorithms*, 26:306–324, 1998.
- [41] 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.
- [42] L. A. J. Müller, K. G. Kugler, A. Dander, A. Graber, and M. Dehmer. QuACN - an R package for analyzing complex biological networks quantitatively. *Bioinformatics*, 27(1):140–141, 2011.
- [43] M. Randić. On characterization of molecular branching. *J. Amer. Chem. Soc.*, 97:6609–6615, 1975.
- [44] M. Randić, P. J. Hansen, and P. C. Jurs. Search for useful graph theoretical invariants of molecular structure. *J.Chem.Inf. Comput.Sci.*, 28:60–68, 1988.
- [45] S. E Stein. NIST, Mass spectral database 98. www.nist.gov/srd/nist1a.htm, 1998. National Institute of Standards and Technology, Gaithersburg, MD, USA.
- [46] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors*. Wiley-VCH, 2002. Weinheim, Germany.
- [47] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *ICML’09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080. ACM, 2009.

- [48] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
- [49] H. Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(17):17–20, 1947.
- [50] Y. Yang. Resistance distances and the global cyclicity index of fullerene graphs. *Digest Journal of Nanomaterials and Biostructures*, 7:593–598, 2012.