OPEN ACCESS Freely available online

PLoS one

Information Indices with High Discriminative Power for Graphs

Matthias Dehmer[1]*, Martin Grabner[1], Kurt Varmuza[2]

1 Institute for Bioinformatics and Translational Research, UMIT, Hall in Tyrol, Austria, **2** Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Vienna, Vienna, Austria

Abstract

In this paper, we evaluate the uniqueness of several information-theoretic measures for graphs based on so-called information functionals and compare the results with other information indices and non-information-theoretic measures such as the well-known Balaban J index. We show that, by employing an information functional based on degree-degree associations, the resulting information index outperforms the Balaban J index tremendously. These results have been obtained by using nearly 12 million exhaustively generated, non-isomorphic and unweighted graphs. Also, we obtain deeper insights on these and other topological descriptors when exploring their uniqueness by using exhaustively generated sets of alkane trees representing connected and acyclic graphs in which the degree of a vertex is at most four.

Citation: Dehmer M, Grabner M, Varmuza K (2012) Information Indices with High Discriminative Power for Graphs. PLoS ONE 7(2): e31214. doi:10.1371/journal.pone.0031214

Editor: Dongxiao Zhu, Wayne State University, United States of America

Received October 14, 2011; **Accepted** January 4, 2012; **Published** February 29, 2012

Funding: Matthias Dehmer, Martin Grabner and Kurt Varmuza thank the Austrian Science Funds for supporting this work (project P22029-N13). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: matthias.dehmer@umit.at

Introduction

To quantify the topology of networks, numerous topological descriptors, which are also often referred to as graph measures or indices, have been developed [1–7]. A property thereof called the *uniqueness*, *discriminative power* or *degeneracy* has been investigated extensively in mathematical chemistry and structure-oriented drug design in the context of characterizing the structure of molecules quantitatively. In general, a descriptor is called *degenerate* if it possesses the same value for more than one graph. In this paper our main task is to examine the extent to which topological indices are degenerate.

We briefly review the most important contributions to tackle this problem, and start with a classical contribution due to Bonchev et al. [8,9]. They proposed the so-called magnitude-based information indices for improving the discriminative power of other classical descriptors for alkane trees [8] and isomers [9]. Alkane trees are connected and acyclic graphs in which the degree of a vertex is at most four [10]. Following this, Raychaudhri et al. [11] analyzed the discriminative power of information-theoretic measures based on distances for chemical graphs containing one ring. Konstantinova et al. [12] explored the uniqueness of various information-theoretic and non-information-theoretic measures by using polycyclic structures representing cata-condensed benzenoid hydrocarbons. As a result, the Balaban J index (see equation 20), the sum of local vertex entropies due to Konstantinova [12,13] and the magnitude-based information indices turned out to be unique for this class of graphs; see [12]. However, note that the sizes of the corresponding sets C_i, denoted by $|C_i|$, were rather small, $2 \leq |C_i| \leq 1681$. Diudea et al. [14] recently explored a novel super-index based on shell matrices and polynomials. By applying this index to the heterogeneous graph database MS2265 [15] containing 2265 non-isomorphic skeleton graphs, inferred from chemical compounds, and to chemical isomers, it turned out that this index does not have any degeneracy [14]. Other results obtained when applying further topological descriptors to chemical graph databases can be also found in [14]. Hu and Xu [16] applied an index using layer matrices and powers of extended adjacency matrices to over two million weighted alkane isomers. The index was unique for all graph classes used [16], but we point out that the developed index is based on using bond types and 3D information.

In order to underpin the practical importance of exploring uniqueness, it seems reasonable that an appropriate graph measure to characterize the structure of networks quantitatively should be able to discriminate graphs properly (e.g., when slightly changing the structure of a network). Note that this problem has already been discussed in the context of complex networks; see [17]. As to applications thereof, Dehmer et al. [15] have already outlined that unique measures can serve as candidates for calculating the identification codes of networks (e.g., chemical structures), which could be used to perform fast structure searches in large databases. Also, such highly discriminating measures representing graph invariants (the measured value is invariant under graph isomorphisms [10]) can be useful to tackle the graph isomorphism problem, because, if the values of two graphs with the same number of vertices are different, they must be non-isomorphic. Hence, such indices could be employed to tackle the graph isomorphism problem in large databases, as the computational complexity of the measures is polynomial. That means instead of performing a thorough isomorphism test which may be computationally costly, highly unique graph measures could be used to filter out non-isomorphic graphs. Note that the time complexity of some of these measures has already been discussed in [15].

The main contribution of this paper is to evaluate the discriminative power of selected topological indices in the context of complex networks, i.e., graphs that are neither regular nor random [18]. We applied several information-theoretic and non-information-theoretic measures, such as the Balaban J index [19], to nearly 12 million exhaustively generated, non-isomorphic and unweighted graphs with the same number of vertices (see 'Numerical results and interpretation'). Importantly, we only use unweighted graphs in this study, as it poses an extra challenge to the underlying descriptors to discriminate such graphs on a large scale. We emphasize that the Balaban J index has often been referred to as one of the most discriminative indices (see e.g. [20]), as it is powerful when applied to several classes of isomers and alkane trees. Our study highlights the limitations of the Balaban J index and other topological descriptors in terms of their ability to discriminate non-isomorphic graphs uniquely.

We prove that one of the information indices due to Dehmer et al. [15,21], which uses the information functional f^{Δ} based on degree-degree associations, outperforms the Balaban J index tremendously when these measures are applied to exhaustively generated graphs. We also employ other information measures for graphs using so-called information functionals that have been developed by Dehmer et al. [15,21]. The discriminative power of some of these information measures and classical ones has already been evaluated in [22] specifically for chemical graphs possessing structural constraints. By contrast, we perform a large-scale study to compare the discriminative power of these information measures by employing three information functionals (see equations 7, 8, and 18) and non-information-theoretic indices such as the Balaban J index using exhaustively generated graphs without structural constraints. The discriminative power by employing these particular information functionals and Balaban J index has not yet been investigated on a large scale.

The results can be interpreted as an attempt to evaluate the uniqueness of quantitative graph measures in the context of complex networks. To the best of our knowledge, very little work has so far been done to tackle this problem. One exception is the work of Kim et al. [17], who evaluated the discriminative power of graph complexity measures that were developed in the context of network physics. As a result, most of the complexity measures proposed in [17] turned out to show little discriminative power.

This paper is organized as follows. In the section 'Topological descriptors' we briefly recall the definitions of the information-theoretic measures due to Dehmer et al. and the other graph measures that we are going to use. The 'Data and software' section describes the datasets and sketches the steps to calculate the topological descriptors. In 'Numerical results and Interpretation', we present and interpret the numerical results when evaluating the discriminative power of the measures. This includes a statistical analysis to investigate the dependence of the uniqueness of the Balaban J index and $I^{\lambda}_{f^{\Delta}}$ on the sample size by using exhaustively generated graphs with 10 vertices. The paper finishes with a 'Summary and conclusion'.

Methods

Topological Descriptors

In this section, we briefly recall the definition of the information measures [4,15,21] that we are going to use in this study. Further, we outline the concept of distance-based descriptors, including the well-known Balaban J index. In summary, Table 1 gives an overview of the descriptors that we use.

Information Indices. To start, we point out that, besides empirical properties of information measures for graphs [1,4,15,21] (such as determining correlations between the measures [1]), mathematical problems (such as proving various upper and lower bounds of the measures) have also been explored; see [23,24]. Note that the correlation ability between two graph measures generally relates to the problem of whether they capture structural information similarly [1,9]. The so-called implicit information inequalities have been investigated extensively in [21,25,26]. Also, the class of graph entropy measures obtained by using certain information functionals based on the metric properties of graphs (such as the neighborhoods of atoms) has been used to solve problems in quantitative structure–activity relationships (QSARs) and quantitative structure–property relationships (QSPRs) [27]. In particular, Dehmer et al. [28] classified the mutagenicity of molecules by using these measures and employing supervised learning techniques.

Table 1. The topological indices used and their symbols.

Index Name	Symbol
Balaban index [19]	J
Balaban-like 1 [36]	U
Balaban-like 2 [36]	X
Bertz index [46]	C_B
Magnitude-based Entropy [8]	I_D
Magnitude-based Entropy [8]	I_D^W
Compactness [7]	C
Complexity Index [2]	B
Vertex Complexity [11]	I_V
Harary index [7]	H
Hyper Distance Path index [7]	D_P
Sum of Vertex Entropies [13]	I_{loc}
Normalized Edge Complexity [2]	E_n
Prod. of Row Sums [7]	PRS
Radial Centric index [1]	$I_{C,R}$
Top. Information Content [31]	I_a
Index of total adjacency [2]	A
Degree Information index [1]	I_δ
Zagreb 1 [7]	Z_1
Zagreb 2 [7]	Z_2
Information index using f^V [15,21]	$I^{\lambda}_{f^V_{lin}}$
Information index using f^V [15,21]	$I^{\lambda}_{f^V_{quad}}$
Information index using f^V [15,21]	$I^{\lambda}_{f^V_{exp}}$
Information index using f^P [15,21]	$I^{\lambda}_{f^P_{lin}}$
Information index using f^P [15,21]	$I^{\lambda}_{f^P_{quad}}$
Information index using f^P [15,21]	$I^{\lambda}_{f^P_{exp}}$
Information index using f^{Δ} [21,34]	$I^{\lambda}_{f^{\Delta}_{exp}}$

doi:10.1371/journal.pone.0031214.t001

Let $G=(V,E)$ be an arbitrary, finite, and unweighted graph; $|V|$ denotes the number of vertices and $|E|$ the number of edges, respectively. Throughout this paper, we use the symbol $|A|$ to express the cardinality (also called the size) of a set A. We denote by $\rho(G)$ the diameter of G; see [29]. The abstract information functionals [21] $f: V \rightarrow_+$ play a critical role when defining information measures on graphs. Based on these functionals, vertex

probabilities [21]

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

have been assigned to each particular vertex of G. This makes the resulting measure independent of determining partitions of graph invariants [1,8,30,31], which might be computationally difficult to obtain. By definition,

$$p^f(v_1)+p^f(v_2)+\cdots+p^f(v_{|V|})=1, \tag{2}$$

and $(p^f(v_1),\ldots,p^f(v_{|V|}))$ therefore forms a probability distribution. Using this approach and recalling Shannon's entropy [32] defined by

$$I=\sum_{i=1}^{n} p_i \log(p_i), \tag{3}$$

the families of information measures

$$I_f(G) := -\sum_{i=1}^{|V|} \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)} \log\left(\frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)}\right), \tag{4}$$

$$I_f^{\lambda}(G) := \lambda\left(\log(|V|)+\sum_{i=1}^{|V|} \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)} \log\left(\frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)}\right)\right) \tag{5}$$

have been developed [4,15,21]. These measures are families of entropic measures representing the structural information content of G. Here $\lambda>0$ is a scaling constant, I_f is the mean entropy of G, and I_f^{λ} its information distance between maximum entropy and I_f.

In our analysis, we define three distinct functionals f^V, f^P, and f^{Δ}, and the relative information measures $I_{f^V}^{\lambda}$, $I_{f^P}^{\lambda}$, and $I_{f^{\Delta}}^{\lambda}$ [4,5,21]. To define f^V, we first define the j-sphere of a vertex $v_i \in V$ by [21]

$$S_j(v_i,G) := \{v \in V | d(v_i,v)=j, j\geq 1\}. \tag{6}$$

$|S_j(v_i,G)|$ are just the j-sphere cardinalities. In general, $d(v_i,v_j)$ is the shortest distance between the vertices $v_i,v_j \in V$; see [33]. Then,

$$f^V(v_i) := c_1|S_1(v_i,G)|+c_2|S_2(v_i,G)|+\cdots+c_{\rho(G)}|S_{\rho(G)}(v_i,G)|, \quad c_k>0,\ 1\leq k\leq \rho(G). \tag{7}$$

To define f^P, the pathlengths for $j=1,2,\ldots,\rho(G)$ of the local information graph $L_G(v_i,j)$ starting from a particular vertex have been used; see [21] for its detailed definition. For example, $P(L_G(v_i,j))$ is the sum of all pathlengths starting from $v_i \in V$ by inducing shortest paths for $j=1,2,\ldots,\rho(G)$. We obtain

$$f^P(v_i) := c_1 l(P(L_G(v_i,1)))+c_2 l(P(L_G(v_i,2))) + \cdots + c_{\rho(G)} l(P(L_G(v_i,\rho(G)))), \quad b_k>0,\ 1\leq k\leq \rho(G). \tag{8}$$

Finally, we define f^{Δ} (see [34]), let G be an undirected and unweighted graph, and set $S_j(v_i,G) := \{v_{a_j},v_{b_j},\ldots,v_{z_j}\}$, $1\leq j\leq \rho(G)$, $1\leq i\leq |V|$. For $v_i \in V$, we define the sets of shortest paths [34]

$$\mathcal{P}_1^j(v_i) := (v_i,v_{a_1}^j,v_{a_2}^j,\ldots,v_{a_j}^j), \tag{9}$$

$$\mathcal{P}_2^j(v_i) := (v_i,v_{b_1}^j,v_{b_2}^j,\ldots,v_{b_j}^j), \tag{10}$$

$$\vdots$$

$$\mathcal{P}_{k_j}^j(v_i) := (v_i,v_{z_1}^j,v_{z_2}^j,\ldots,v_{z_j}^j), \tag{11}$$

and the corresponding degree sequences [34]

$$s_1^j(v_i) := (\delta(v_i),\delta(v_{a_1}^j),\delta(v_{a_2}^j),\ldots,\delta(v_{a_j}^j)), \tag{12}$$

$$s_2^j(v_i) := (\delta(v_i),\delta(v_{b_1}^j),\delta(v_{b_2}^j),\ldots,\delta(v_{b_j}^j)), \tag{13}$$

$$\vdots$$

$$s_{k_j}^j(v_i) := (\delta(v_i),\delta(v_{z_1}^j),\delta(v_{z_2}^j),\ldots,\delta(v_{z_j}^j)). \tag{14}$$

The quantities [34]

$$\Delta^G(v_i,1)=|\delta(v_i)-\delta(v_{a_1}^1)|+\cdots+|\delta(v_i)-\delta(v_{z_1}^1)|, \tag{15}$$

$$\Delta^G(v_i,2)=|\delta(v_i)-\delta(v_{a_1}^2)|+\cdots+|\delta(v_i)-\delta(v_{z_1}^2)| + \cdots+|\delta(v_{z_1}^2)-\delta(v_{z_2}^2)|, \tag{16}$$

$$\vdots$$

$$\Delta^G(v_i,\rho(G))=|\delta(v_i)-\delta(v_{a_1}^{\rho(G)})|+\cdots+|\delta(v_{a_{\rho(G)-1}}^{\rho(G)})-\delta(v_{a_{\rho(G)}}^{\rho(G)})| + |\delta(v_i)-\delta(v_{z_1}^{\rho(G)})|+\cdots+|\delta(v_{z_{\rho(G)-1}}^{\rho(G)})-\delta(v_{z_{\rho(G)}}^{\rho(G)})| \tag{17}$$

have been used to define the information functional f^{Δ}; see equation 18. As we employ the differences $|\delta(v)-\delta(u)|$, the resulting graph entropies $I_{f^{\Delta}}$ and $I_{f^{\Delta}}^{\lambda}$ have been called *degree–degree association indices*; see [34]. Now, f^{Δ} has been defined by [34]

$$f^{\Delta}(v_i) := \alpha^{c_1\Delta^G(v_i,1)+c_2\Delta^G(v_i,2)+\cdots+c_{\rho(G)}\Delta^G(v_i,\rho(G))}, \quad c_k>0,\ 1\leq k\leq \rho(G),\ \alpha>0. \tag{18}$$

We see that f^{Δ} is well defined for any $\alpha>0$. Since f^V, f^P and f^{Δ} as well as the resulting entropies are parametric, we need to choose the

coefficients c_i for weighting the structural differences or characteristics of a graph. Note that the c_k must be chosen such that at least two coefficients c_i,c_j are distinct. This includes the parameter settings, e.g.,

$$c_1 > c_2 > \cdots > c_\rho \quad or \quad c_1 < c_2 < \cdots < c_\rho, \tag{19}$$

which have already been used in [15]. Other configurations of the c_i have also been investigated to determine the structural complexity of chemical structures meaningfully [15].

Distance-Based Topological Descriptors. Numerous topological descriptors have been explored by employing distances in a graph [7,19,29]. Seminal work was done by Skorobogatov and Dobrynin [29], who developed a theory on the metric properties of graphs. Also, several distance-based graph measures have been developed and analyzed where these indices have shown that distances in graphs capture significant information when applied in QSAR/QSPR; see [1,7,11,19,27].

We recall the definition of the Balaban J index [7,19] in detail as we place emphasis on comparing its discriminative power with $I^{\lambda}_{f^V}$, $I^{\lambda}_{f^P}$, and $I^{\lambda}_{f^\Delta}$ on a large scale by using exhaustively generated graphs. The names and symbols of the remaining descriptors used in this study can be found in Table 1. For their formal definitions, see [1,2,7,27].

Now, we define the distance matrix [35] of a graph G as $DS=(d(v_i,v_j))_{i,j}$. For each vertex $v_i \in V$, DS_i denotes the distance sum (row or column sum) obtained by adding the entries in the corresponding row or column of the distance matrix DS. In addition, $\mu := |E|+1-|V|$ is the cyclomatic number [36]. Then, the Balaban J index is defined by [19]

$$J(G) := \frac{|E|}{\mu+1} \sum_{(v_i,v_j)\in E} [DS_i DS_j]^{-1/2}. \tag{20}$$

Results

Data and Software

Let us now state the definitions and generation procedure of the graphs for performing our analysis.

Definition 1 *N_i is the set of all exhaustively generated non-isomorphic and connected graphs with i vertices.*

Practically, these sets have been generated by using the program geng from the Nauty package [37]. In this study we use the classes $N_5,\ldots,N_{10}$ and obtain their cardinalities as follows: $|N_5|=21$, $|N_6|=112$, $|N_7|=853$, $|N_8|=11117$, $|N_9|=261080$, and $|N_{10}|=11716571$. These numbers are in accordance with the results due to McKay [37,38].

Definition 2 *C_i is the set of all exhaustively generated non-isomorphic alkane trees graphs with i vertices.*

The chemical structures represented by alkane trees with a carbon backbone have been generated with Molgen [39]. In particular, we generated the classes $C_{19},\ldots,C_{22}$; their cardinalities are $|C_{19}|=148284$, $|C_{20}|=366319$, $|C_{21}|=910726$, and $|C_{22}|=2278658$.

Then for both classes (see Definitions 1 and 2), the structure information has been converted into the graphNEL format to calculate the descriptors in R [40] by employing the QuACN package [41]. This package contains R functions of over a hundred topological descriptors.

Numerical Results and Interpretation

In this section, we present the numerical results when evaluating the discriminative power of the information indices, Balaban J index and other topological descriptors. Results on exhaustively generated graphs are summarized in Tables 2 and 3, while those on alkane trees are given in Table 5. In total, we evaluated the discriminative power of 27 graph measures.

Evaluation of the Discriminative Power Using Exhaustively Generated Graphs. To interpret the numerical results, we start by considering Table 3 and observe that the sensitivity values due to Konstantinova [12], $S=(|\mathcal{G}|-\mathrm{ndv})/|\mathcal{G}|$, for Balaban J decreases with increasing number of vertices; see also the 'Statistical analysis' section. Throughout this paper, ndv (non-distinguishable values) stands for the number of non-isomorphic graphs whose values cannot be distinguished by a particular index [12]. For example, by considering the class N_8, 61.6623% of the graphs could be distinguished (i.e., have unique values) by the Balaban J index. For N_{10}, only 20.5633% out of

Table 2. N_5, N_6 and N_7 are exhaustive sets of non-isomorphic and connected graphs. $|N_5|=21$, $|N_6|=112$ and $|N_7|=853$.

	N_5		N_6		N_7	
Index	**ndv**	***S***	**ndv**	***S***	**ndv**	***S***
J	0	1,000000	10	0,910714	155	0,818288
U	0	1,000000	10	0,910714	155	0,818288
X	0	1,000000	10	0,910714	155	0,818288
C_B	20	0,047619	111	0,008929	852	0,001172
I_D	15	0,285714	100	0,107143	826	0,031653
I_D^W	14	0,333333	94	0,160714	811	0,049238
C	16	0,238095	108	0,035714	847	0,007034
B	2	0,904762	34	0,696429	486	0,430246
I_V	10	0,523810	91	0,187500	797	0,065651
H	14	0,333333	100	0,107143	828	0,029308
D_P	14	0,333333	101	0,098214	837	0,018757
I_{loc}	2	0,904762	34	0,696429	450	0,472450
E_n	19	0,095238	110	0,017857	851	0,002345
PRS	2	0,904762	38	0,660714	486	0,430246
$I_{C,R}$	20	0,047619	111	0,008929	852	0,001172
I_a	20	0,047619	111	0,008929	852	0,001172
A	19	0,095238	110	0,017857	851	0,002345
I_δ	20	0,047619	111	0,008929	852	0,001172
Z_1	19	0,095238	110	0,017857	851	0,002345
Z_2	0	1,000000	37	0,669643	750	0,120750
$I^{\lambda}_{f^V_{lin}}$	4	0,809524	37	0,669643	485	0,431419
$I^{\lambda}_{f^V_{quad}}$	4	0,809524	37	0,669643	452	0,470106
$I^{\lambda}_{f^V_{exp}}$	4	0,809524	37	0,669643	454	0,467761
$I^{\lambda}_{f^P_{lin}}$	9	0,571429	38	0,660714	312	0,634232
$I^{\lambda}_{f^P_{quad}}$	2	0,904762	23	0,794643	97	0,886284
$I^{\lambda}_{f^P_{exp}}$	2	0,904762	5	0,955357	7	0,991794
$I^{\lambda}_{f^\Delta_{exp}}$	6	0,714286	16	0,857143	34	0,960141

doi:10.1371/journal.pone.0031214.t002

Table 3. Exhaustive sets of non-isomorphic graphs. $|N_8| = 11117$, $|N_9| = 261080$, $|N_{10}| = 11716571$.

	N_8		N_9		N_{10}	
Index	**ndv**	***S***	**ndv**	***S***	**ndv**	***S***
J	4262	0,616623	156674	0,399900	9307263	0,205633
U	4093	0,631825	148132	0,432618	8812811	0,247834
X	4093	0,631825	148132	0,432618	8812810	0,247834
C_B	11116	0,000090	261079	0,000004	11716570	0,000000
I_D	11070	0,004228	260971	0,000417	11716339	0,000020
I_D^W	11014	0,009265	260803	0,001061	11715858	0,000061
C	11110	0,000630	261072	0,000031	11716564	0,000001
B	8384	0,245840	237199	0,091470	11472695	0,020815
I_V	10958	0,014302	260650	0,001647	11715029	0,000132
H	11076	0,003688	261018	0,000237	11716455	0,000010
D_P	11100	0,001529	261054	0,000100	11716541	0,000003
I_{loc}	8305	0,252946	235233	0,099000	11395248	0,027425
E_n	11115	0,000180	261078	0,000008	11716569	0,000000
PRS	9376	0,156607	252262	0,033775	11672850	0,003732
$I_{C,R}$	11116	0,000090	261079	0,000004	11716570	0,000000
I_a	11116	0,000090	261079	0,000004	11716570	0,000000
A	11115	0,000180	261078	0,000008	11716569	0,000000
I_δ	11116	0,000090	261079	0,000004	11716570	0,000000
Z_1	11115	0,000180	261078	0,000008	11716569	0,000000
Z_2	10996	0,010884	260931	0,000571	11716379	0,000016
$I^{\lambda}_{f^V_{lin}}$	9165	0,175587	249439	0,044588	11640381	0,006503
$I^{\lambda}_{f^V_{quad}}$	8300	0,253396	235044	0,099724	11385762	0,028234
$I^{\lambda}_{f^V_{exp}}$	8300	0,253396	235055	0,099682	11385730	0,028237
$I^{\lambda}_{f^P_{lin}}$	4989	0,551228	158391	0,393324	9479777	0,190909
$I^{\lambda}_{f^P_{quad}}$	1699	0,847171	58196	0,777095	4243499	0,637821
$I^{\lambda}_{f^P_{exp}}$	478	0,957003	27017	0,896518	2619898	0,776394
$I^{\lambda}_{f^\Delta_{exp}}$	385	0,965368	6016	0,976957	609204	0,948005

doi:10.1371/journal.pone.0031214.t003

almost 12 million exhaustively generated non-isomorphic graphs could be distinguished by J. But we can see in Table 3 that the information indices using the information functional approach [4,15,21] sketched in the 'Information indices' section can discriminate our graphs comparatively well. In particular, $I^{\lambda}_{f^\Delta}$, with an exponential weighting scheme

$$c_1 := \rho(G),\ c_2 := \rho(G)e^{-1},\ \ldots,\ c_{\rho(G)} := \rho(G)e^{-\rho(G)+1}, \quad (21)$$

denoted by $I^{\lambda}_{f^\Delta_{\exp}}$, discriminates 94.8005% out of almost 12 million exhaustively generated graphs successfully. In view of the large number and complexity of the graphs (see $|N_8|$, $|N_9|$ and $|N_{10}|$), the uniqueness of $I^{\lambda}_{f^\Delta}$ is striking. Observe that, for all weighting schemes [15], i.e., lin, quad, and exp, $I^{\lambda}_{f^V}$ is much less discriminative. We realize that the underlying information functional f is crucial for reaching uniqueness of the information index. Also, we can clearly see that the uniqueness of other indices shown in Table 3 is quite low. We see that the Balaban U and X indices are among the best out of the set of known measures that we have chosen to perform this study.

Interestingly, the situation is somewhat the opposite when considering Table 2. Namely, for N_5 and N_6, the discriminative power of the Balaban J index is higher than by using some of the information measures based on the information functional approach (e.g., $I^{\lambda}_{f^V_{lin}}$ and $I^{\lambda}_{f^P_{lin}}$). Also, we see that the underlying weighting scheme for the coefficients matters a lot, because $I^{\lambda}_{f^P_{\exp}}$ has a higher discriminative power than the Balaban J index for N_6 and N_7. In summary, we hypothesize that the Balaban J index performs well if the cardinality of the underlying graph set and the order of the involved graphs is rather small. By using a statistical approach, we will verify this hypothesis in the 'Statistical analysis' section. Let us give another example to shed light on the degeneracy of the measures when applying them to graphs $\in N_{10}$, see Figure 1 and Table 4. Figure 1 shows four sample graphs $\in N_{10}$ where G_3 and G_4 are structurally quite similar in the following sense. If we remove the edge $\{2,10\}$ in G_3 and the edge $\{6,10\}$ in G_4, the resulting graphs are isomorphic. From Table 4, we see that these graphs can only be fully distinguished by the degree-degree association index. Evaluating the Balaban J index on these graphs gives two degenerate graphs namely G_1 and G_2. In contrast to this, I_{loc} due to Konstantinova can not discriminate G_3 and G_4. Finally, we observe that I_a can not

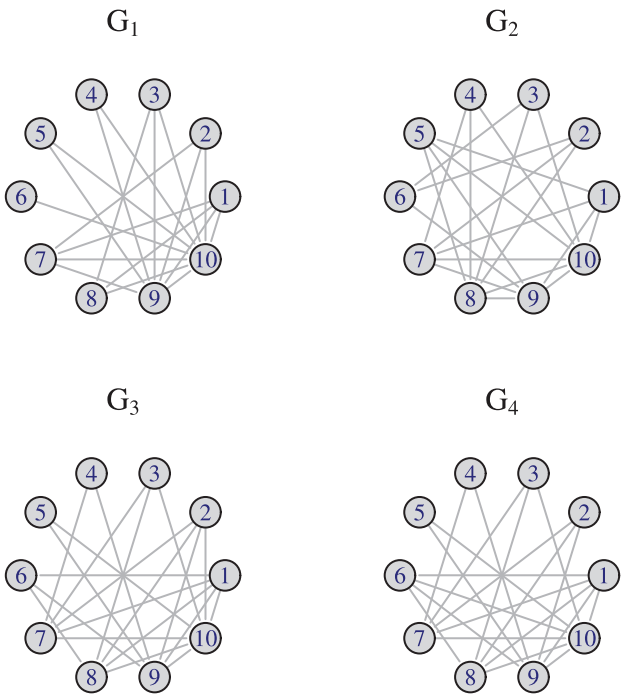

Figure 1. Four example graphs $G_i \in N_{10}$.
doi:10.1371/journal.pone.0031214.g001

discriminate any of the four example graphs. This implies that every measure captures structural information differently and, hence, its discriminative power can differ dramatically because of

- the underlying paradigm to define a graph measure, e.g., information-theoretic vs. non-information-theoretic indices or partition-based vs. non-partition-based
- the underlying graph invariant to define a measure, e.g., degrees or distances or several graph invariants etc.

A comparison of the measures with others (e.g., see Table 3) is critical, as the measures rely on different concepts (e.g., information-theoretic vs.non-information-theoretic indices). In the following, we give plausible reasons why the measures using the information functional approach often capture structural information of exhaustively generated graphs more uniquely and significantly than other information measures for graphs that are based on determining partitions of graph invariants. This can also be underpinned by the numerical results; see Tables 2 and 3. Examples of the latter measures are the magnitude-based information indices I_D and I_D^W due to Bonchev et al. [8], the degree information index I_δ [1] and the topological information content of a graph I_a [31,42].

Table 4. Index values for the four example graphs depicted in Figure 1.

	$I^{\lambda}_{f^{\Delta}_{exp}}$	J	I_{loc}	I_a
G_1	0.0002695	2.639475	31.16882	3.121928
G_2	0.8801102	2.633647	30.90633	3.321928
G_3	0.2076738	2.564776	30.92375	3.321928
G_4	0.0017872	2.564776	30.92375	3.321928

doi:10.1371/journal.pone.0031214.t004

To construct classical partition-based measures of a graph G, we start with a graph invariant X and induce a partitioning according to an equivalence criterion. This results in the equivalence classes $X_1,\ldots,X_k$ being obtained. The mean entropy is then given by

$$I(G) = -\sum_{i=1}^{k} \frac{|X_i|}{|X|} \log\left(\frac{|X_i|}{|X|}\right). \tag{22}$$

The process of inducing the partitionings might be the reason for obtaining non-unique indices, as many structurally different graphs could possess the same or similar partitionings when using a certain equivalence criterion, e.g., vertex degree equality [1] or topologically equivalent vertices [31,42].

In order to derive information measures using the information functional approach, we assign a probability value (see equation 1) to each individual vertex in a graph by using a certain information functional f capturing its structural information. Examples thereof are equations 7 and 18. That means the information measures given by equations 4 and 5 can be understood as a cumulation of local quantities representing the vertex probabilities. Clearly, each such quantity captures a certain percentage rate of the structure of G. As the numerical results show, these measures conserve structural information more properly than the partition-based ones and result in highly discriminating measures for several graph classes. Note that other classical descriptors (see Tables 2 and 3), such as the Harary index, Randic′ index [43,44] and the complexity index B etc., rely on the simple derivation of structural quantities (e.g., distances or degrees) to obtain a single numerical value characterizing the complexity the graph. Consequently, their discriminative power is very low; see Tables 2 and 3.

When evaluating the uniqueness (see ndv or S values) of $I^{\lambda}_{f^{V}_{\exp}}$ and $I^{\lambda}_{f^{\Delta}_{\exp}}$ (see Table 3), we observe that the difference between the resulting values is tremendous. Note that the graphs of N_8,N_9, and N_{10} contain cycles. A plausible reason for this is given in Figure 2.

We see on the left-hand side that the j-sphere cardinalities are rather small if j goes to $\rho(G)$ and, hence, their contribution to the value of the particular functional for v_i is small too. Also, there is not much variation between the j-sphere cardinalities. This could be a reason that the resulting probability values

$$p^{f^V}(v_i) = \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)},$$

are quite similar to each other and, thus, this has a direct influence on the resulting value of the information index and on its uniqueness. In contrast, the right-hand side of Figure 2 shows that the values of $\Delta^G(v_i,j)$ are more diverse and, in particular, those values when j goes to $\rho(G)$ are larger than the j-sphere cardinalities. This might be a plausible reason why the corresponding vertex probability values are more different and, hence, the resulting entropies as well. As Tables 2 and 3 show, we again emphasize that the discriminative power of an index clearly depends on the underlying graph class.

Evaluation of the Discriminative Power by Using Chemical Graphs. Here we evaluate the uniqueness of the Balaban J index, the information measures using the information functional approach, and the remaining topological descriptors shown in Table 1 by also using chemical graphs. Table 5 depicts the numerical results when applying the measures to chemical alkane trees representing the skeletal graphs. The number of

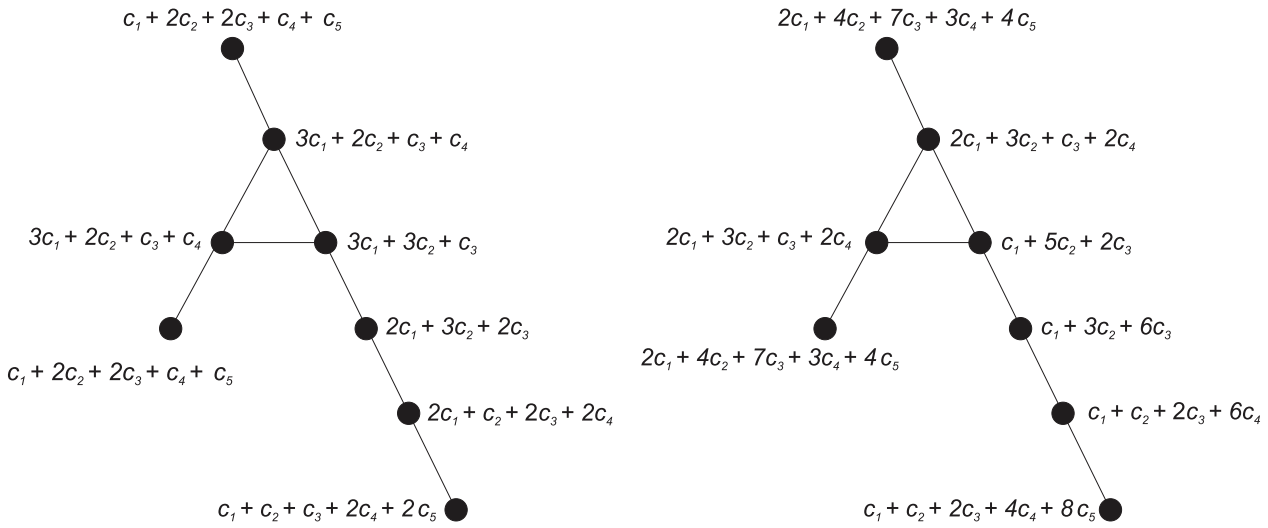

Figure 2. Left: A cyclic graph and its values of f^V for each vertex. Right: Values of f^Δ for each vertex for the same graph.
doi:10.1371/journal.pone.0031214.g002

vertices ranges from 19 to 22. We see again that the discriminative power of the Balaban J index decreases when the number of graphs and vertices increase. The Balaban-like indices possess high discriminative power for all four graph classes. Also, we observe that the sum of the local vertex entropies (I_{loc}) due to Konstantinova [13,45] has high uniqueness. Interestingly, it is as good as $I^{\lambda}_{f^V_{\exp}}$ and $I^{\lambda}_{f^P_{\exp}}$. It can be easily shown that, for trees, the information indices using f^V and f^P have equal discriminative power. In particular, $I^{\lambda}_{f^V_{\exp}}$, $I^{\lambda}_{f^P_{\exp}}$ and the just mentioned indices clearly outperform the Balaban J index by using the chemical alkane trees.

Finally, the numerical results show again that the discriminative power of a structural index strongly depends on the underlying graph class. See, for instance, the results when comparing the uniqueness of $I^{\lambda}_{f^{\Delta}_{\exp}}$ for the alkane trees and exhaustively generated graphs (see Table 3).

Descriptive Statistical Analysis. In order to provide further evidence for stability of the uniqueness of $I^{\lambda}_{f^{\Delta}}$ by using exhaustively generated graphs, we perform a statistical analysis by using boxplots. The graph class to perform the study is N_{10}. It is clear that, for computational reasons, the statistical analysis cannot be performed by using the entire set N_{10}. Hence, we choose subsets of N_{10} whose sizes are called sample sizes. Also, we perform the boxplot analysis for Balaban J as well, and present the resulting plots to investigate the dependence between uniqueness and sample size; see Figure 3. Concretely, 100 samples of 1100, 3300, 11 000, 33 000, 100 000, and 333 000 randomly chosen graphs out of N_{10} have been analyzed by standard R boxplot routines. That means the medians have been calculated and plotted, with the first and third quantiles as hinges. The whiskers represent the calculated borders of the 95% confidence interval.

As we can see in Figure 3 the uniqueness values are not dispersed for a given sample size, but they depend on the sample size. Further, we observe that the uniqueness of the Balaban J index is not stable when the sample size is varied. In general, we call a measure I unstable if there is a strong dependency between the uniqueness of I and the sample size to perform the statistical analysis. In contrast, I is stable if there is only a very little dependency between the uniqueness of I and the sample size.

We see from the boxplot that the uniqueness decreases if the sample size increases. Based on our intuition, it seems reasonable that, the smaller the sample size, the better is the discriminative power of the measure under consideration. Thus $I^{\lambda}_{f^{\Delta}}$ possesses a non-trivial property, namely a very high discriminative power for exhaustively generated graphs that is almost independent of sample size. By using the above stated definition, we see that $I^{\lambda}_{f^{\Delta}}$ is stable on N_{10} as the uniqueness is constantly high and does not depend much on the sample size. We see from Table 3 that $I^{\lambda}_{f^{\Delta}_{\exp}}$ is the only topological descriptor possessing this property. Other topological measures, and particularly the Balaban J index, have the trivializing property that, for exhaustively generated graphs, the uniqueness is only reasonable for small sets of graphs.

Hence some of the entropy measures using the information functional approach could be applied successfully for discriminating sets of large complex networks as well. Keep in mind that in fact such classes of exhaustively generated complex networks possess huge cardinalities. Note that the cardinality of the exhaustively generated non-isomorphic graphs with 10 vertices is already greater than 11 million. As we conclude from this statistical analysis, $I^{\lambda}_{f^{\Delta}_{\exp}}$ possesses the stability property that is necessary to achieve feasible results when applied to sets of large complex networks.

Summary and Conclusion

In this paper, we have dealt with the problem of evaluating the discriminative power of topological graph measures by using exhaustively generated, non-isomorphic graphs without vertex and edge weights. We have made an attempt to translate topological indices into the field of complex networks when evaluating their uniqueness. We found that one of the information measures for graphs using the information functional based on degree–degree associations outperformed the Balaban J index tremendously. Also, by using the graph class N_{10}, we found that the uniqueness of the Balaban J index is quite sensitive to varying sample size when performing the statistical analysis; see 'Statistical analysis' section. In particular, the uniqueness of the Balaban J index deteriorated when increasing the sample size. This makes Balaban J in particular non-feasible for discriminating complex networks structurally as they are multicyclic, do not have structural constraints, and the cardinality of an underlying set of such networks is huge. This property was also observed by using other topological indices shown in Table 1. The numerical results when

Table 5. Chemical alkane trees $T=(V,E)$ with $|V|=19,\ldots,22$. $|C_{19}|=148284$, $|C_{20}|=366319$, $|C_{21}|=910726$, $|C_{22}|=2278658$.

	C_{19}		C_{20}		C_{21}		C_{22}	
Index	**ndv**	S	**ndv**	S	**ndv**	S	**ndv**	S
J	5967	0,959760	44800	0,877702	45703	0,949817	306911	0,865311
U	0	1,000000	12	0,999967	4	0,999996	82	0,999964
X	0	1,000000	12	0,999967	4	0,999996	82	0,999964
C_B	148278	0,000040	366312	0,000019	910718	0,000009	2278645	0,000006
I_D	68030	0,541218	171655	0,531406	452442	0,503207	1140578	0,499452
I_D^W	39731	0,732061	97815	0,732979	277238	0,695586	702776	0,691583
C	148267	0,000115	366289	0,000082	910713	0,000014	2278626	0,000014
B	5959	0,959814	44752	0,877833	45667	0,949857	306469	0,865505
I_V	104790	0,293316	279826	0,236114	730474	0,197921	1942075	0,147711
H	125290	0,155067	319121	0,128844	813614	0,106631	2081153	0,086676
D_P	147946	0,002279	365914	0,001106	910290	0,000479	2278165	0,000216
I_{loc}	0	1,000000	12	0,999967	4	0,999996	84	0,999963
E_n	148283	0,000007	366318	0,000003	910725	0,000001	2278657	0,000000
PRS	5967	0,959760	44810	0,877675	45701	0,949819	306953	0,865292
$I_{C,R}$	148283	0,000007	366318	0,000003	910725	0,000001	2278656	0,000001
I_a	148278	0,000040	366312	0,000019	910718	0,000009	2278645	0,000006
A	148283	0,000007	366318	0,000003	910725	0,000001	2278657	0,000000
I_δ	148283	0,000007	366318	0,000003	910725	0,000001	2278657	0,000000
Z_1	148283	0,000007	366318	0,000003	910725	0,000001	2278657	0,000000
Z_2	148282	0,000013	366317	0,000005	910724	0,000002	2278656	0,000001
$I^{\lambda}_{f^V_{lin}}$	5006	0,966241	37820	0,896757	39210	0,956946	263231	0,884480
$I^{\lambda}_{f^V_{quad}}$	42	0,999717	268	0,999268	324	0,999644	1752	0,999231
$I^{\lambda}_{f^V_{exp}}$	0	1,000000	12	0,999967	4	0,999996	84	0,999963
$I^{\lambda}_{f^P_{lin}}$	5006	0,966241	37820	0,896757	39210	0,956946	263231	0,884480
$I^{\lambda}_{f^P_{quad}}$	42	0,999717	268	0,999268	324	0,999644	1752	0,999231
$I^{\lambda}_{f^P_{exp}}$	0	1,000000	12	0,999967	4	0,999996	84	0,999963
$I^{\lambda}_{f^{\Delta}_{exp}}$	67176	0,546977	196124	0,464609	544432	0,402200	39396	0,982711

doi:10.1371/journal.pone.0031214.t005

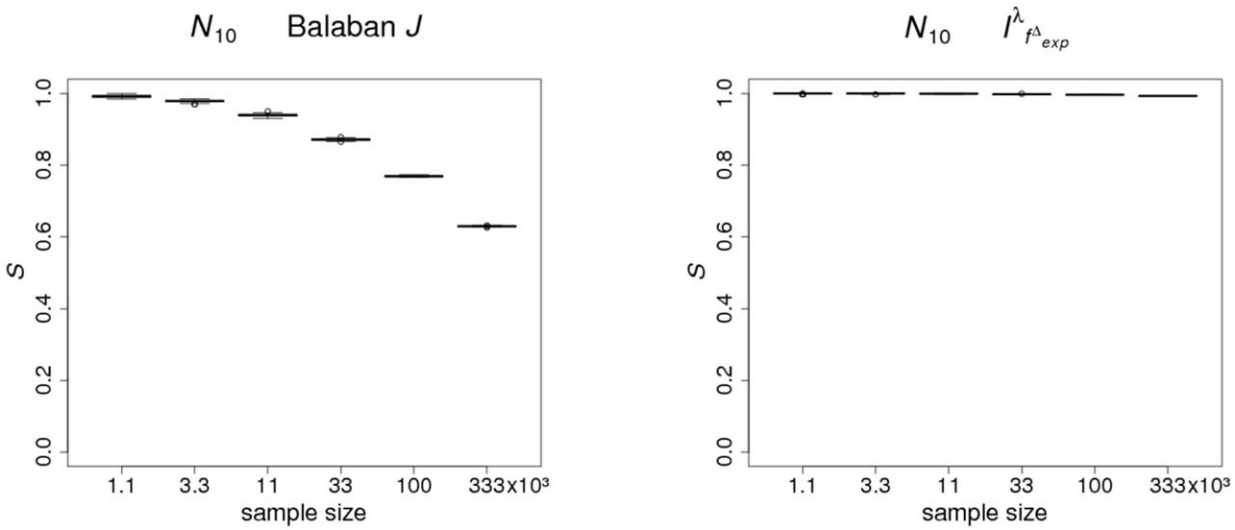

Figure 3. Boxplots to investigate the dependency of the uniqueness of Balaban J and $I^{\lambda}_{f^{\Delta}}$ from the sample size by using exhaustively generated graphs with ten vertices.
doi:10.1371/journal.pone.0031214.g003

using exhaustively generated graphs and alkane trees can be found in Tables 2, 3, and 5.

Altogether, this study clearly shows the limitations of topological indices and restrictions when applying them on a large scale. A topological index can be unique for a particular graph class but it fails when applying the measure to another class. In this sense, it is far from trivial that we obtained an index (see the definition of $I_{f^A}^{\lambda}$) that turned out to be highly discriminating for exhaustively generated graph classes. Note that the underlying graphs do not possess structural constraints.

As to future work, we will evaluate further topological indices on a large scale to obtain deeper theoretical insights. From such an analysis, one can also learn how the measures capture structural information. This relates to better understanding of their structural interpretation. We are convinced that these developments could also trigger future developments positively when developing and investigating topological graph measures in the context of complex networks.

Author Contributions

Analyzed the data: MD MG KV. Wrote the paper: MD MG KV.

References

1. Bonchev D (1983) Information Theoretic Indices for Characterization of Chemical Structures. Research Studies Press, Chichester.
2. Bonchev D, Rouvray DH (2005) Complexity in Chemistry, Biology, and Ecology. Mathematical and Computational Chemistry. Springer. New York, NY, USA.
3. da F Costa L, Rodrigues F, Travieso G (2007) Characterization of complex networks: A survey of measurements. Advances in Physics 56: 167–242.
4. Dehmer M, Mowshowitz A (2011) A history of graph entropy measures. Information Sciences 1: 57–78.
5. Emmert-Streib F, Dehmer M (2007) Information theoretic measures of UHG graphs with low computational complexity. Applied Mathematics and Computation 190: 1783–1794.
6. Mehler A, Weiß P, Lücking A (2010) A network model of interpersonal alignment. Entropy 12: 1440–1483.
7. Todeschini R, Consonni V, Mannhold R (2002) Handbook of Molecular Descriptors. Wiley-VCH. Weinheim, Germany.
8. Bonchev D, Trinajstić N (1977) Information theory, distance matrix and molecular branching. J Chem Phys 67: 4517–4533.
9. Bonchev D, Mekenyan O, Trinajstić N (1981) Isomer discrimination by topological information approach. J Comp Chem 2: 127–148.
10. Trinajstić N (1992) Chemical Graph Theory. CRC Press. Boca Raton, FL, USA.
11. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984) Discrimination of isomeric structures using information theoretic topological indices. Journal of Computational Chemistry 5: 581–588.
12. Konstantinova EV (1996) The discrimination ability of some topological and information distance indices for graphs of unbranched hexagonal systems. J Chem Inf Comput Sci 36: 54–57.
13. Konstantinova EV, Paleev AA (1990) Sensitivity of topological indices of polycyclic graphs. Vychisl Sistemy 136: 38–48.
14. Diudea MV, Ilić A, Varmuza K, Dehmer M (2011) Network analysis using a novel highly discriminating topological index. Complexity 16: 32–39.
15. Dehmer M, Varmuza K, Borgert S, Emmert-Streib F (2009) On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. J Chem InfModel 49: 1655–1663.
16. Xu CYHL (1996) On highly discriminating molecular topological index. J Chem Inf Comput Sci 36: 82–90.
17. Kim J, Wilhelm T (2008) What is a complex graph? Physica A 387: 2637–2652.
18. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks. From Biological Networks to the Internet and WWW. Oxford University Press.
19. Balaban AT (1982) Highly discriminating distance-based topological index. Chem Phys Lett 89: 399–404.
20. Vukičević D, Balaban AT (2005) On the degeneracy of topological index J. Internet Electronic Journal of Molecular Design 4: 491–500.
21. Dehmer M (2008) Information processing in complex networks: Graph entropy and information functionals. Appl Math Comput 201: 82–94.
22. Dehmer M, Barbarini N, Varmuza K, Graber A (2009) A large scale analysis of informationtheoretic network complexity measures using chemical structures. PLoS ONE 4: e8057.
23. Li X, Gutman I (2006) Mathematical Aspects of Randić-Type Molecular Structure Descriptors. Mathematical Chemistry Monographs. University of Kragujevac and Faculty of Science Kragujevac.
24. Zhou B (2008) Bounds on the balaban index. Croatica Chemica Acta 81: 319–323.
25. Dehmer M, Borgert S, Emmert-Streib F (2008) Entropy bounds for molecular hierarchical networks. PLoS ONE 3: e3079.
26. Dehmer M, Borgert S, Bonchev D (2008) Information inequalities for graphs. Symmetry: Culture and Science Symmetry in Nanostructures (Special issue edited by M Diudea) 19: 269–284.
27. Devillers J, Balaban AT (1999) Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach Science Publishers. Amsterdam, The Netherlands.
28. Dehmer M, Barbarini N, Varmuza K, Graber A (2010) Novel topological descriptors for analyzing biological networks. BMC Structural Biology 10.
29. Skorobogatov VA, Dobrynin AA (1988) Metrical analysis of graphs. Commun Math Comp Chem 23: 105–155.
30. Bonchev D (2009) Information theoretic measures of complexity. In: Meyers R, ed. Encyclopedia of Complexity and System Science, Springer, volume 5. pp 4820–4838.
31. Mowshowitz A (1968) Entropy and the complexity of the graphs I: An index of the relative complexity of a graph. Bull Math Biophys 30: 175–204.
32. Shannon CE, Weaver W (1949) The Mathematical Theory of Communication. University of Illinois Press.
33. Dijkstra EW (1959) A note on two problems in connection with graphs. Numerische Math 1: 269–271.
34. Dehmer M, Emmert-Streib F, Tsoy Y, Varmuza K (2011) Quantifying structural complexity of graphs: Information measures in mathematical chemistry. In: Putz M, ed. Quantum Frontiers of Atoms and Molecules, Nova Publishing. pp 479–498.
35. Harary F (1969) Graph Theory. Addison Wesley Publishing Company. Reading, MA, USA.
36. Balaban AT, Balaban TS (1991) New vertex invariants and topological indices of chemical graphs based on information on distances. J Math Chem 8: 383–397.
37. McKay BD (2010). Nauty. http://cs.anu.edu.au/_bdm/nauty/.
38. McKay BD (1998) Isomorph-free exhaustive generation. Journal of Algorithms 26: 306–324.
39. (2000). Molgen isomer generator software. www.molgen.de. Institute of Mathematics II, University of Bayreuth, Germany.
40. (2011). R, software, a language and environment for statistical computing. www.r-project.org. R Development Core Team, Foundation for Statistical Computing, Vienna, Austria.
41. Müller LAJ, Kugler KG, Dander A, Graber A, Dehmer M (2010) QuACN - an R package for analyzing complex biological networks quantitatively. Bioinformatics. pp 140–141.
42. Rashevsky N (1955) Life, information theory, and topology. Bull Math Biophys 17: 229–235.
43. Randić M (1975) On characterization of molecular branching. J Amer Chem Soc 97: 6609–6615.
44. Wiener H (1947) Structural determination of paraffin boiling points. Journal of the American Chemical Society 69: 17–20.
45. Konstantinova EV, Skorobogatov VA, Vidyuk MV (2002) Applications of information theory in chemical graph theory. Indian Journal of Chemistry 42: 1227–1240.
46. Bertz SH (1981) The first general index of molecular complexity. Journal of the American Chemical Society 103: 3241–3243.