Information Indices with High Discriminative Power for Graphs

Matthias Dehmer1, Martin Grabner1, Kurt Varmuza2

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.

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 polymeric 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 $C_j$, denoted by $|C_j|$, were rather small, $2 \leq |C_j| \leq 1681$. 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 $f^A$ 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 $I_{Z_{l}}$ 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-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 $I_{Z_{l}}$ on the sample size by using exhaustively generated graphs with 10 vertices. The paper finishes with a ‘Summary and conclusion’.

**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.

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] $f: V \rightarrow \mathbb{R}$ play a critical role when defining information measures on graphs. Based on these functionals, vertex
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,

$$p'(v_i) = \frac{f(v_i)}{\sum_{j=1}^{[V]} f(v_j)} \quad (1)$$

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

$$I = \sum_{i=1}^{[V]} p_i \log(p_i), \quad (3)$$

the families of information measures

$$I_f(G) = -\sum_{i=1}^{[V]} \frac{f(v_i)}{\sum_{j=1}^{[V]} f(v_j)} \log \left( \frac{f(v_i)}{\sum_{j=1}^{[V]} f(v_j)} \right) \quad (4)$$

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

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 $I_f^j$ its information distance between maximum entropy and $I_f$.

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

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

$|S_j(v_i,G)|$ 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,

$$f^V(v_i) = c_1 |S_1(v_i,G)| + c_2 |S_2(v_i,G)| + \cdots + c_k |S_k(v_i,G)|, \quad c_k > 0, 1 \leq k \leq \rho(G) \quad (7)$$

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, $P(L_G(v_i,1))$ 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

$$f^P(v_i) = c_1 L(P(L_G(v_i,1))) + c_2 L(P(L_G(v_i,2))) + \cdots + c_k L(P(L_G(v_i,\rho(G)))), \quad (8)$$

$$b_k > 0, 1 \leq k \leq \rho(G).$$

Finally, we define $f^A$, let $G$ be an undirected and unweighted graph, and set $S_j(v_i,G) = \{ v_j, v_{j_1}, \ldots, v_{j_z} \}$,

$$1 \leq j \leq \rho(G), 1 \leq i \leq |V|.$$ For each $v_i \in V$, we define the sets of shortest paths [34]

$$P_i^j(v_i) = \{ v_i, v_{i_1}, v_{i_2}, \ldots, v_{i_j} \}. \quad (9)$$

$$P_i^2(v_i) = \{ v_i, v_{i_1}, v_{i_2} \}, \quad (10)$$

$$\vdots$$

$$P_i^j(v_i) = \{ v_i, v_{i_1}, v_{i_2}, \ldots, v_{i_j} \}. \quad (11)$$

and the corresponding degree sequences [34]

$$d_i^j(v_i) = (\delta(v_i), \delta(v_{i_1}), \delta(v_{i_2}), \ldots, \delta(v_{i_j})) \quad (12)$$

$$d_i^2(v_i) = (\delta(v_i), \delta(v_{i_1}), \delta(v_{i_2}), \ldots, \delta(v_{i_j})) \quad (13)$$

$$\vdots$$

$$d_i^j(v_i) = (\delta(v_i), \delta(v_{i_1}), \delta(v_{i_2}), \ldots, \delta(v_{i_j})) \quad (14)$$

The quantities [34]

$$\Delta^G(v_i,1) = |\delta(v_i) - \delta(v_{i_1})| + \cdots + |\delta(v_i) - \delta(v_{i_j})| \quad (15)$$

$$\Delta^G(v_i,2) = |\delta(v_i) - \delta(v_{i_1}^2)| + \cdots + |\delta(v_i) - \delta(v_{i_j}^2)| \quad (16)$$

$$\vdots$$

$$\Delta^G(v_i,\rho(G)) = |\delta(v_i) - \delta(v_{i_1}^\rho(G))| + \cdots + |\delta(v_i) - \delta(v_{i_j}^\rho(G))| \quad (17)$$

have been used to define the information functional $f^A$; see equation 18. As we employ the differences $|\delta(v_i) - \delta(u_i)|$, the resulting graph entropies $I_A^j$ and $I_A^j$ have been called degree–degree association indices; see [34]. Now, $f^A$ has been defined by [34]

$$f^A(v_i) = c_1^A \Delta^G(v_i,1) + c_2^A \Delta^G(v_i,2) + \cdots + c_k^A \Delta^G(v_i,\rho(G)) \quad (18)$$

$$c_k > 0, 1 \leq k \leq \rho(G), z > 0.$$
coefficients $c_i$ for weighting the structural differences or characteristics of a graph. Note that the $c_i$ must be chosen such that at least two coefficients $c_i, c_j$ are distinct. This includes the parameter settings, e.g.,

$$c_1 > c_2 > \cdots > c_p \quad \text{or} \quad c_1 < c_2 < \cdots < c_p,$$

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 $I_P$, $I_C$, and $I_{PC}$ 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 $DS = (d(v_i, v_j))_{ij}$. 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]

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

### 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 sets $N_5, \ldots, N_{10}$ and their cardinalities as follows: $|N_5| = 21$, $|N_6| = 112$, $|N_7| = 853$, $|N_8| = 11117$, $|N_9| = 261080$, and $|N_{10}| = 1716571$. 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], $S = (|G| - ndv)/|G|$, 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 |
| $J$ | 0 | 1,000000 | 10 | 0,910714 | 155 | 0,818288 |
| $X$ | 0 | 1,000000 | 10 | 0,910714 | 155 | 0,818288 |
| $C_P$ | 20 | 0,047619 | 111 | 0,008929 | 852 | 0,001172 |
| $I_P$ | 15 | 0,285714 | 100 | 0,107143 | 826 | 0,031653 |
| $I_{PC}$ | 14 | 0,333333 | 94 | 0,160714 | 811 | 0,049238 |
| $C_2$ | 16 | 0,238095 | 108 | 0,035714 | 847 | 0,007034 |
| $B$ | 2 | 0,904762 | 34 | 0,696429 | 486 | 0,430246 |
| $I_P$ | 10 | 0,523810 | 91 | 0,187500 | 797 | 0,065651 |
| $H$ | 14 | 0,333333 | 100 | 0,107143 | 828 | 0,029308 |
| $I_P$ | 14 | 0,333333 | 101 | 0,098214 | 837 | 0,018757 |
| $I_{PC}$ | 2 | 0,904762 | 34 | 0,696429 | 450 | 0,472450 |
| $E_1$ | 19 | 0,952388 | 110 | 0,017857 | 851 | 0,003245 |
| $I_{PR}$ | 2 | 0,904762 | 38 | 0,660714 | 486 | 0,430246 |
| $I_{PR}$ | 20 | 0,047619 | 111 | 0,008929 | 852 | 0,001172 |
| $I_{PC}$ | 20 | 0,047619 | 111 | 0,008929 | 852 | 0,001172 |
| $I_{PC}$ | 19 | 0,952388 | 110 | 0,017857 | 851 | 0,003245 |
| $Z_2$ | 0 | 1,000000 | 37 | 0,669643 | 750 | 0,120750 |
| $I_{PC}$ | 4 | 0,809524 | 37 | 0,669643 | 485 | 0,431419 |
| $I_{PC}$ | 4 | 0,809524 | 37 | 0,669643 | 452 | 0,470106 |
| $I_{PC}$ | 4 | 0,809524 | 37 | 0,669643 | 454 | 0,467761 |
| $I_{PC}$ | 9 | 0,571429 | 38 | 0,660714 | 312 | 0,634232 |
| $I_{PC}$ | 2 | 0,904762 | 23 | 0,794628 | 97 | 0,886284 |
| $I_{PC}$ | 2 | 0,904762 | 5 | 0,953537 | 7 | 0,991794 |
| $I_{PC}$ | 6 | 0,714286 | 16 | 0,857143 | 34 | 0,960141 |

doi:10.1371/journal.pone.0031214.t002
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, $I_{2,h}$, 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 $I_{2,h}^{3,\exp}$, discriminates 94.8005% out of almost 12 million exhaustively generated graphs successfully. In view of the large number and complexity of the graphs [see $|N_9|$, $|N_0|$ and $|N_{10}|$], the uniqueness of $I_{2,h}^{3,\exp}$ is striking. Observe that, for all weighting schemes [15], i.e., lin, quad, and exp, $I_{2,h}$ 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 ($e.g., I_{2,h}^{3,\exp}$ and $I_{2,h}^{3,\lin}$). Also, we see that the underlying weighting scheme for the coefficients matters a lot, because $I_{2,h}^{3,\exp}$ 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 $G_1$ and $G_2$ are structurally quite similar in the following sense. If we remove the edge $\{2,10\}$ in $G_1$ and the edge $\{6,10\}$ in $G_2$, 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 rise to degenerate graphs namely $G_1$ and $G_2$. In contrast to this, $I_{2,w}$ due to Konstantinova can not discriminate $G_1$ and $G_2$. Finally, we observe that $I_{d}$ can not

| Table 3. Exhaustive sets of non-isomorphic graphs. $|N_9| = 11117$, $|N_0| = 261080$, $|N_{10}| = 11716571$. |
|---|---|---|---|---|---|
| Index | $N_9$ | $N_0$ | $N_{10}$ |
| | 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 |
| $H^B_{l,h}$ | 11014 | 0.009265 | 260803 | 0.001061 | 11715858 | 0.000061 |
| $C$ | 11110 | 0.000630 | 261072 | 0.000031 | 11716564 | 0.000001 |
| $R$ | 8384 | 0.245840 | 237199 | 0.091470 | 11472695 | 0.020815 |
| $I_V$ | 10958 | 0.014302 | 260650 | 0.001647 | 11715029 | 0.000132 |
| $H$ | 11076 | 0.003688 | 260108 | 0.000237 | 11716455 | 0.000000 |
| $I_D$ | 11100 | 0.005159 | 261054 | 0.001010 | 11716541 | 0.000003 |
| $I_{lin}$ | 8305 | 0.252946 | 235233 | 0.099900 | 11395248 | 0.027425 |
| $I_{exp}$ | 11115 | 0.000180 | 261078 | 0.000008 | 11716569 | 0.000000 |
| $PRS$ | 9376 | 0.156607 | 252262 | 0.033775 | 11672850 | 0.003732 |
| $I_{C.A}$ | 11116 | 0.000090 | 261079 | 0.000004 | 11716570 | 0.000000 |
| $I_{a}$ | 11116 | 0.000090 | 261079 | 0.000004 | 11716570 | 0.000000 |
| $A$ | 11115 | 0.001800 | 261078 | 0.000080 | 11716569 | 0.000000 |
| $I_{s}$ | 11116 | 0.000090 | 261079 | 0.000080 | 11716569 | 0.000000 |
| $Z_1$ | 11115 | 0.001800 | 261078 | 0.000080 | 11716569 | 0.000000 |
| $Z_2$ | 10996 | 0.010884 | 260931 | 0.000570 | 11716379 | 0.000016 |
| $I_{lin}^{3,\exp}$ | 9165 | 0.175587 | 249439 | 0.044588 | 11640381 | 0.006503 |
| $I_{exp}^{3,\exp}$ | 8300 | 0.253396 | 235044 | 0.099724 | 11385762 | 0.028234 |
| $I_{lin}^{3,\lin}$ | 8300 | 0.253396 | 235055 | 0.099682 | 11385730 | 0.028237 |
| $I_{exp}^{3,\lin}$ | 4989 | 0.551228 | 158391 | 0.393324 | 9479777 | 0.190909 |
| $I_{lin}^{3,\quad}$ | 1699 | 0.847171 | 58196 | 0.777095 | 4243499 | 0.637821 |
| $I_{exp}^{3,\quad}$ | 478 | 0.957003 | 27017 | 0.896518 | 2619898 | 0.776394 |
| $I_{d}$ | 385 | 0.965368 | 6016 | 0.976957 | 609204 | 0.948005 |
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

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

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, Randić 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 of the graph. Consequently, their discriminative power is very low; see Tables 2 and 3.

When evaluating the uniqueness (see ndv or $S$ values) of $I_D$ and $I_d$ (see Table 3), we observe that the difference between the resulting values is tremendous. Note that the graphs of $N_4, N_6, N_{10}$ 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

$$P^{fv}(v_i) = \frac{f^{fv}(v_i)}{\sum_{j=1}^{\rho(G)} f^{fv}(v_j)},$$

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 $\Delta^d(v_i,j)$ 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
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 \( I_{l_{exp}} \) due to Konstantinova [13,45] has high uniqueness. Interestingly, it is as good as \( I_{P_{exp}} \) and \( I_{l_{exp}} \). It can be easily shown that, for trees, the information indices using \( f^T \) and \( f^P \) have equal discriminative power. In particular, \( I_{l_{exp}}, I_{P_{exp}} \), 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 \( I_{l_{exp}} \) 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 \( I_{l_{exp}} \) 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 \( I_{l_{exp}} \) 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 \( I_{l_{exp}} \) 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 \( I_{l_{exp}} \) 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, \( I_{l_{exp}} \) 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 alkanes trees \( T = (V,E) \) with \( |V| = 19, \ldots, 22 \). \( |C_{19}| = 148284, |C_{20}| = 366319, |C_{21}| = 910726, |C_{22}| = 2278658. \)

| Index | \( C_{19} \) | \( C_{20} \) | \( C_{21} \) | \( C_{22} \) |
|-------|-------------|-------------|-------------|-------------|
| \( J \) | 5967        | 0.959760    | 44800       | 0.877702    | 45703       | 0.949817    | 306911      | 0.865311    |
| \( U \) | 0           | 1.000000    | 12          | 0.999967    | 4           | 0.999999    | 82          | 0.999964    |
| \( X \) | 0           | 1.000000    | 12          | 0.999967    | 4           | 0.999999    | 82          | 0.999964    |
| \( C_{19} \) | 148278     | 0.000040    | 366312      | 0.000019    | 910718      | 0.000009    | 2278645     | 0.000006    |
| \( I_{10} \) | 68030      | 0.541218    | 171655      | 0.531406    | 452442      | 0.503207    | 1140578     | 0.499452    |
| \( \mu_{10}^{w} \) | 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    |
| \( I_{10} \) | 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    |
| \( \sigma \) | 147946     | 0.002279    | 365914      | 0.001106    | 910290      | 0.000479    | 2279165     | 0.000216    |
| \( I_{10} \) | 0          | 1.000000    | 12          | 0.999967    | 4           | 0.999999    | 84          | 0.999963    |
| \( E_{u} \) | 148283     | 0.000007    | 366318      | 0.000003    | 910725      | 0.000001    | 2278626     | 0.000000    |
| \( PRS \) | 5967       | 0.959760    | 44810       | 0.877833    | 45701       | 0.949819    | 306953      | 0.865292    |
| \( I_{10} \) | 148283     | 0.000007    | 366318      | 0.000003    | 910725      | 0.000001    | 2278656     | 0.000001    |
| \( I_{10} \) | 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    |
| \( I_{10} \) | 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    |
| \( \mu_{10}^{v} \) | 5006       | 0.966241    | 37820       | 0.896757    | 39210       | 0.956946    | 263231      | 0.884480    |
| \( \mu_{10}^{u} \) | 42         | 0.999717    | 268         | 0.999268    | 324         | 0.999644    | 1752        | 0.999231    |
| \( \mu_{10}^{w} \) | 5006       | 1.000000    | 12          | 0.999967    | 4           | 0.999999    | 84          | 0.999963    |
| \( \mu_{10}^{v} \) | 5006       | 0.966241    | 37820       | 0.896757    | 39210       | 0.956946    | 263231      | 0.884480    |
| \( \mu_{10}^{u} \) | 42         | 0.999717    | 268         | 0.999268    | 324         | 0.999644    | 1752        | 0.999231    |
| \( \mu_{10}^{w} \) | 0          | 1.000000    | 12          | 0.999967    | 4           | 0.999999    | 84          | 0.999963    |
| \( \mu_{10}^{v} \) | 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 \( I_{10}^{v} \) 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 3.

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 $I_{2}(G)$) 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.