A Large Scale Analysis of Information-Theoretic Network Complexity Measures Using Chemical Structures

Matthias Dehmer[1]*, Nicola Barbarini[2], Kurt Varmuza[3], Armin Graber[1]

[1] Institute for Bioinformatics and Translational Research, UMIT, Hall in Tyrol, Austria, [2] Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy, [3] Institute of Chemical Engineering, Laboratory for Chemometrics, Vienna University of Technology, Vienna, Austria

Abstract

This paper aims to investigate information-theoretic network complexity measures which have already been intensely used in mathematical- and medicinal chemistry including drug design. Numerous such measures have been developed so far but many of them lack a meaningful interpretation, e.g., we want to examine which kind of structural information they detect. Therefore, our main contribution is to shed light on the relatedness between some selected information measures for graphs by performing a large scale analysis using chemical networks. Starting from several sets containing real and synthetic chemical structures represented by graphs, we study the relatedness between a classical (partition-based) complexity measure called the topological information content of a graph and some others inferred by a different paradigm leading to partition-independent measures. Moreover, we evaluate the uniqueness of network complexity measures numerically. Generally, a high uniqueness is an important and desirable property when designing novel topological descriptors having the potential to be applied to large chemical databases.

Citation: Dehmer M, Barbarini N, Varmuza K, Graber A (2009) A Large Scale Analysis of Information-Theoretic Network Complexity Measures Using Chemical Structures. PLoS ONE 4(12): e8057. doi:10.1371/journal.pone.0008057

Editor: Christopher L. Douglas, University of California, Berkeley, United States of America

Received July 30, 2009; **Accepted** October 22, 2009; **Published** December 15, 2009

Funding: This work was supported by the COMET Center ONCOTYROL and funded by the Federal Ministry for Transport Innovation and Technology (BMVIT) and the Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth (BMWA/BMWFJ), the Tiroler Zukunftsstiftung (TZS) and the State of Styria represented by the Styrian Business Promotion Agency (SFG) [and supported by the University for Health Sciences, Medical Informatics and Technology and BIOCRATES Life Sciences AG]. Also, funding from the FIRB ITALBIONET Project is gratefully acknowledged. 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

The problem to quantify the complexity of a network appears in various scientific disciplines [1–7] and has been a challenging research topic of ongoing interest for several decades [8]. This problem first appeared when studying the complexity of biological and chemical systems, e.g., battery cells or living systems [9–12] using information-theoretic measures [13] (in this paper, we use the words "measure", "index", "descriptor" synonymously when referring to topological graph complexity measures). Directly afterwards, the idea of applying entropy measures to network-based systems finally emerged as a new branch in mathematical complexity science. An important problem within this area deals with determining the so-called structural information content [8,12,14–19] of a network. Finally, it turned out that the developed information indices for measuring the information content of a graph have been of substantial impact when solving QSPR (Quantitative structure-property relationship)/QSAR (Quantitative structure-activity relationship) problems in mathematical chemistry and drug design [1,2,20–25]. Correspondingly, such measures have been widely used to predict biological activities as well as toxicological and physico-chemical properties of molecules using chemical datasets, see, e.g., [1,20,23–26]. More precisely, most powerful and generally applicable for theses approaches are empirical multivariate models $y=f(\mathbf{x})$, with y being a chemical or a physical property (P) or a biological activity (A), and vector $\mathbf{x}$ consisting of a series of numerical molecular descriptors describing the molecular structure. For modeling biological activities also (measured or computed) physical properties are used. Some of the already mentioned information-theoretic complexity measures which are well-established in mathematical chemistry will be defined in the next section.

Before sketching the aims of our paper, we start with a brief review about classical and more recent approaches to measure the complexity of networks. However, for performing the numerical results, we mainly restrict our analysis to information-theoretic measures which are based on Shannon's entropy [13] and which have already been applied in the context of mathematical chemistry [2,21] and drug design [1,20,23].

In general, it seems clear that *complexity* and, even, *structural complexity* is generally not uniquely defined because it is in the eye of a beholder [27]. Consequently, it is often not clear which structural features of a graph in question should be taken into account. For instance, to use complexity measures within mathematical chemistry, some of their desirable features were stated in [3]. Now, we start outlining the most known classical approaches and then turn to more recently developed approaches for detecting network complexity. Beside the already mentioned information-based measures [1,2,8,20–26,28], the complexity of a network was also defined by using boolean functions approaches [6,8,29,30]. For example, Constantine [29] defined the complexity of a graph to be the number of its containing spanning trees. Jukna [30] determined graph complexity as the minimum number of union and intersection operations required to obtain the whole set of its edges starting from star graphs. Finally, the so-called combinatorial complexity of a network was developed by Minoli [6]. The key property of such a descriptor is that it must be a monotonically increasing function of the factors which contribute to the complexity of a network, e.g., number of vertices and edges, vertex degrees (branching [3]), multiple edges, cycles, loops, and labels [3]. Another crucial definition of complexity (algorithmic information) that is different compared to the mentioned ones was given by Kolmogorov [31]. Based on appropriate string encodings of graphs, bounds to estimate the Kolmogorov-complexity of labeled and unlabeled graphs were obtained in [32]. However, this kind of network complexity measures are difficult to apply in general because of computational reasons [32]. In order to briefly review more recently developed approaches, we start by mentioning some quantities for structurally characterizing networks [33,34] which emerged from complex network theory [33,35–37]:

- Size of the giant connected component [33,38].
- Degree distributions $P(i)$ [33,38–41].
- Exponent of degree distributions [33], i.e., it holds $P(i) \sim i^{-\gamma}$.
- Total number of vertices and edges [33,34,40,42,43].
- Path-based quantities [33,40,42,44].
- Distance-based quantities, e.g., j-spheres, average distances, eccentricity, diameter and radius [33,40,42,44].
- Degree, degree statistics and edge density [33,40,42,44].
- Clustering coefficient, modularity and network motifs [45–48].
- Eigenvector measures [40,49,50].

Further, various measures have been developed to characterize the complexity of networks where many of the recent ones were summarized by Kim et al. [51] and da Costa et al. [44]. In particular, information-theoretic complexity measures for general graphs have been investigated in [51–54]. For instance, starting from directed networks, the information measure called Medium Articulation was defined which is maximized for exactly the medium number of links [53]. Properties thereof were examined in [54]. Another entropy-based measure called Offdiagonal complexity (OdC) was contributed by Claussen [52]. This graph complexity measure is based on determining the entropy of the so-called offdiagonal elements of the vertex-vertex link correlation matrix [51,52]. Similar entropy measures can be also found in [44,55]. We already mentioned that the number of spanning trees might also serve as graph complexity measure, see, e.g. [29]. As a further attempt, Kim et al. [51] developed a more sophisticated approach by calculating a quantity for each edge that takes the number of spanning trees of the graph and the number of spanning trees of the corresponding one-edge-deleted subgraph into account. By using these entities which were called sensitivities, an entropic measure was defined and interpreted as a spanning tree sensitivity complexity of a network. Another important class of network complexity measures is based on determining subgraphs of a network [51,56]. More precisely, the concrete idea is as follows: The more different subgraphs a network contains, the more complex is the underlying network [51]. Here, "different" means that non-isomorphic graphs are considered, however, the graph isomorphism problem is known to be computationally costly, see, e.g. [57,58]. Thus, Kim et al. [51] proposed approximations for decide graph isomorphism and ended up with several subgraph-based graph complexity measures which can be found in [51]. Further, methods based on characterizing subgraph relationships were developed in [56]. To finalize our review on general graph complexity measures, we mention two recently developed approaches [59,60]. In [59], measures were proposed capturing features around each vertex to identify singular vertices. As an interesting result, they found that the obtained singular motifs had unique functional roles in the considered network [59]. A statistical method was defined in [60] to detect network regularity interpreted as simplicity. Finally, starting from a set of measurements and by applying PCA analysis, they found simple regions in the networks under consideration [60]. Interestingly, we want to point out that these two approaches are particularly interesting for investigating biological networks (but not limited to). Especially, the latter method takes incompleteness or noise during the network construction into account [60]. However, the chemical graphs we will use in our paper are deterministically inferrable and not erroneous (measurement errors). This is the reason why we restrict our analysis to information-theoretic measures for globally quantifying the information content of chemical structures where the probability distribution is deterministically inferrable from structural features (e.g., orbits and j-spheres) of the graphs in question.

In this paper, we investigate information-theoretic network complexity measures which are particularly relevant for enhancing empirical QSAR/QSPR models [23]. As we have already expressed, a variety of graph measures have been used so far to characterize the so-called molecular complexity [3,6,61,62]. However, many of such complexity measures lack a meaningful interpretation. Thus, as the major contribution of our paper, we put the emphasis on examining interrelations between information-theoretic network complexity measures often used in mathematical chemistry, that is, we shed light on the problem which kind of structural information the measures detect when applied to chemical graphs.

To tackle this problem, we select a few measures from two different paradigms for inferring such indices: The so-called topological information content [19] (see Equation (4) of a graph and information measures (see Equation (23)) based on using special information functionals [63–65]. The former represents a classical *partition-based* measure that relies on symmetry with respect to topologically equivalent vertices having the same degrees. The latter is a *partition-independent* information measure that is based on using a special information functional capturing structural features of the networks. In order to perform this study, we evaluate these measures numerically by using several large datasets containing real and synthetic chemical graphs. To our best knowledge, such a large scale analysis involving the classical topological information content has not been done so far. Note that in this study, we only consider skeletons of the chemical structures, that is, all atoms are equal and all bonds are equal. Another problem we want to address in this paper is to investigate the uniqueness of complexity measures. This relates to examine their discrimination power, that means, their ability to discriminate non-isomorphic graphs as unique as possible. For this, we also use the mentioned databases - real and synthetic chemical structures - and calculate a special sensitivity measure [66]. Besides evaluating the uniqueness of the information-theoretic measure introduced in the next section, we will calculate the sensitivity values of the entropic measure Offdiagonal complexity and the graph index (Cr is a non-information-theoretic graph complexity measure) Cr, see [51]. Finally, our research addresses the challenging problem of investigating the capability of information-theoretic network descriptors for meaningfully capturing structural features of graphs.

Methods

This section aims to present the information-theoretic topological descriptors we want to investigate in this paper. In the following, we briefly shed light on the two main procedures (resulting in partition-based and partition-independent measures) to infer information-theoretic complexity measures for characterizing chemical network structures. Afterwards, we express their concrete definitions for performing our numerical analysis.

Information-Theoretic Network Complexity Measures

Applying information-theoretic methods for exploring complex networks is a still challenging and ongoing problem [7–9,12,14,15,19,55,67,68]. As mentioned in the introduction, this research area has its origin in biology and mathematical chemistry [8,9,12,19]. Historically seen, Trucco [12] and Rashevsky [19] were the first who developed information measures to analyze complex biological and chemical systems. Later, Mowshowitz [15–18] further developed this approach and proved important mathematical properties thereof.

More precisely, Trucco [12] and Rashevsky [19] defined entropy measures for graphs which were interpreted as the structural information content of a graph; the original information measure due to Rashevsky [19] is called the so-called topological information of a graph in question, see Equation (4). So far, the just mentioned information measures representing the entropy of the underlying graph topology have been widely used for measuring the structural complexity of graphs [3,15–18,21,27,55,69]. The basic principle to infer these measures is as follows: Let $G=(V,E)$ be a graph. By starting from an arbitrary graph invariant X of G and an equivalence criterion α, one obtains a partitioning of X where the partitions are denoted by $X_1,\ldots,X_k$. In order to infer probabilities for each obtained partition, the entities $p_i:=\frac{|X_i|}{|X|}$ can be used because it obviously holds

$$\sum_{i=1}^{k} p_i = 1. \tag{1}$$

Thus $P(G):=(p_1,\ldots,p_k)$ represents a finite probability distribution of G. Now, applying Shannon's entropy formulas [13] leads to the classical graph entropies [8]:

$$I(G,\alpha)=|X|\log(|X|)-\sum_{i=1}^{k}|X_i|\log(|X_i|), \tag{2}$$

$$\bar{I}(G,\alpha)=-\sum_{i=1}^{k}\frac{|X_i|}{|X|}\log\left(\frac{|X_i|}{|X|}\right). \tag{3}$$

Equation (2) is the total information content of G, whereas Equation (3) represents its mean information [2,70]. We want to point out that the just explained procedure yields to partition-based information measures for determining the structural complexity of networks. For example, Mowshowitz [15] obtained such a measure based on algebraic equivalence criteria, e.g., graph automorphisms and graph colorings [15,57]. But it is known that the problem of determining graph automorphisms is equivalent to check whether two graphs are isomorphic [71]. Moreover, the computation of the chromatic number of undirected graphs to infer chromatic decompositions was proven to be NP-complete [58]. Hence, one can expect that the computational complexity of the underlying algorithms for calculating these measures are for arbitrary graphs very costly. After this seminal work [15–19], the outlined principle of inducing vertex partitions was generalized by associating a weighted finite probability distribution to a network, see [8]. This generalization led to numerous information-theoretic graph complexity measures by applying equivalence criteria like vertex degrees, distances to chemical graphs etc. [2,8,21].

Now, we give a sketch of the second procedure for inferring graph entropy measures that results in obtaining partition-independent measures [63–65]. The main idea is as follows: Instead of inducing vertex partitions to obtain probabilities for subsets of vertices, we assign a probability value to every vertex in a graph. This has been done by means of so-called information functionals [64,65] (note that concrete information functionals will be defined in the next section) which capture structural features of a graph and here represent positive mappings which are assumed to be monotonous, see, e.g., [63]. A notable feature of this procedure is that we avoid the problem of determining vertex partitions associated with an equivalence relation that can be often computationally expensive.

As follows, we start with the definition of some concrete partition-based entropy measures to be applied to real and synthetic chemical structures. Note that in this paper, we only evaluate the mean information contents. For the sake of simplicity, we write $I(G)$ instead of $\bar{I}(G)$.

Definition 1. *Let $G=(V,E)$ be a graph.*

$$I_{orb}(G):=-\sum_{i=1}^{k}\frac{|N_i|}{|V|}\log\left(\frac{|N_i|}{|V|}\right), \tag{4}$$

is called topological information content of G. Here, $|N_i|$ denotes the number of topologically equivalent vertices in the i-th vertex orbit of G where k is the number of different orbits.

Remark 1. *Let $G=(V,E)$ be a graph. We recall the definition [2] for two vertices $v,u\in V$ being topologically equivalent: For each i-th neighboring vertex of v there exists an i-th neighboring vertex of u which possesses the same degree. A vertex orbit is a set of vertices that only contains topologically equivalent vertices.*

Definition 2. *Let $G=(V,E)$ be a graph.*

$$I_D(G):=-\frac{1}{|V|}\log\left(\frac{1}{|V|}\right)-\sum_{i=1}^{\rho(G)}\frac{2k_i}{|V|^2}\log\left(\frac{2k_i}{|V|^2}\right), \tag{5}$$

$$I_D^W(G):=-\sum_{i=1}^{\rho(G)}\frac{ik_i}{W}\log\left(\frac{i}{W}\right), \tag{6}$$

where

$$W(G):=\frac{1}{2}\sum_{i=1}^{|V|}\sum_{j=1}^{|V|}d(v_i,v_j). \tag{7}$$

W is called the Wiener index [72] and $d(v_i,v_j)$ denotes the shortest distance between $v_i,v_j\in V$. I_D and I_D^W are so-called magnitude-based information indices, see [69]. It is assumed that the distance of a value i in the distance matrix D appears $2k_i$ times. $\rho(G)$ stands for the diameter of a graph G.

Definition 3. *Let $G=(V,E)$ be a graph.*

$$I_U(G):=\frac{|E|}{\mu+1}\sum_{(v_i,v_j)\in E}[u(v_i)u(v_j)]^{-\frac{1}{2}}, \tag{8}$$

$$I_W(G) := \frac{|E|}{\mu+1}\sum_{(v_i,v_j)\in E}[w(v_i)w(v_j)]^{-\frac{1}{2}}, \tag{9}$$

where

$$u(v_i) := -\sum_{j=1}^{\sigma(v_i)}\frac{jg_j}{d(v_i)}\log\left(\frac{j}{d(v_i)}\right), \tag{10}$$

$$w(v_i) := -w(v_i) = d(v_i)\log(d(v_i)) - u(v_i), \tag{11}$$

$$d(v_i) := \sum_{j=1}^{|V|} d(v_i, v_j) = \sum_{j=1}^{\sigma(v_i)} jg_j. \tag{12}$$

See [28]. g_j equals the number of vertices having distance j starting from $v_i \in V$. Also, g_j equals the corresponding j-sphere cardinality. $\sigma(v) := \max_{u\in V} d(u,v)$ is the eccentricity of $v \in V$. $\mu := |E|+1-|V|$ denotes the cyclomatic number, see [28].

Definition 4. *Let $G=(V,E)$ be a graph.*

$$I_{g_\mu}(v_i) := -\sum_{j=1}^{|V|}\frac{g_\mu^j(v_i)}{\sum_{j=1}^{|V|} g_\mu^j(v_i)}\log\left(\frac{g_\mu^j(v_i)}{\sum_{j=1}^{|V|} g_\mu^j(v_i)}\right), \tag{13}$$

where

$$g_1^j(v_i) := d(v_i,v_j),\, 1 \le i \le |V|, \tag{14}$$

$$g_2^j(v_i) := c_j d(v_i, v_j),\, 1 \le i \le |V|,\, c_i > 0. \tag{15}$$

$I_{g_\mu}(v_i)$ is a local vertex entropy [66]. Finally, the entropy of G can be defined by

$$I_{g_\mu}(G) := \frac{\sum_{i=1}^{|V|} I_{g_\mu}(v_i)}{|V|}. \tag{16}$$

In particular, we define special information measures for characterizing graphs by choosing concrete coefficients [73].

Definition 5. *Let $G=(V,E)$ be a graph. We define*

$$I_{loc}^1(G) := I_{g_1}(G) = \frac{\sum_{i=1}^{|V|} I_{g_1}(v_i)}{|V|}, \tag{17}$$

$$I_{loc}^2(G) := I_{g_2}(G) = \frac{\sum_{i=1}^{|V|} I_{g_2}(v_i)}{|V|}, \tag{18}$$

where

$$c_1 := \rho(G), c_2 := \rho(G)-1, \ldots, c_{\rho(G)} := 1. \tag{19}$$

Finally,

$$I_{loc}^3(G) := I_{g_2}(G), \tag{20}$$

where

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

To finalize this section, we now express the definitions of some partition-independent entropy measures for graphs introduced by Dehmer et al. [63–65]. Mathematical properties and applications thereof can be found, e.g., in [64,65].

Definition 6. *Let $G=(V,E)$ be a graph. The following partition-independent entropy measures based on a special information functional were defined as [63,65]*

$$I_{f^V}(G) := -\sum_{i=1}^{|V|} p^V(v_i)\log(p^V(v_i)), \tag{22}$$

$$I_{f^V}^\lambda(G) := \lambda\left(\log(|V|) + \sum_{i=1}^{|V|} p^V(v_i)\log(p^V(v_i))\right), \tag{23}$$

where $\lambda > 0$ is a scaling constant.

$$p^V(v_i) := \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}, \tag{24}$$

are vertex probabilities. The special information functional f^V was defined as [63]

$$\begin{aligned} f^V(v_i) &:= c_1|S_1(v_i,G)| + c_2|S_2(v_i,G)| + \cdots + c_{\rho(G)}|S_{\rho(G)}(v_i,G)|, \\ &\quad c_k > 0,\, 1 \le k \le \rho(G). \end{aligned} \tag{25}$$

Here, $S_j(v_i, G)$ denotes the j-sphere [65] of a vertex v_i, that is, the set of vertices having shortest distance j starting from $v_j \in V$. c_k are positive coefficients for emphasizing certain structural of a graph, e.g., high vertex degrees, also see, [63,65].

Remark 2. *To perform the numerical calculations in this paper, we set $\lambda = 1000$.*

Definition 7. *Let $G=(V,E)$ be a graph. The measure $I_{f^V}^\lambda$ becomes to $I_{f_{lin}^V}^\lambda$ by choosing the coefficients c_k according to Equation (19), i.e., linearly decreasing. Correspondingly, $I_{f^V}^\lambda$ becomes to $I_{f_{exp}^V}^\lambda$ when choosing the coefficients c_k according to Equation (21), i.e., exponentially decreasing.*

In the following, we briefly comment on the computational complexity of the discussed information measures without giving proofs. Obviously, the measures whose definitions are based on calculating matrices can be often computed in polynomial time (e.g., square, cubic etc.). For instance, it has been proven [74] that the fastest general algorithm to compute the Wiener index is $O(|V||E|)$. Applying W to trees, its computation even only requires time complexity $O(|V|)$. To calculate I_{orb}, the automorphism group of the corresponding graph has to be formally determined. However, it is well known that this procedure is computationally extensive for arbitrary graphs [71]. Hence, this measure is rather not suitable to calculate the information content of large networks. If $G=(V,E)$ is an undirected and connected graph, we showed in [65] that the computation of f^V requires time complexity $O(|V|^2)$. By applying a shortest path algorithm $|V|$ times, it easily follows $I_{f^V}(G)$ has time complexity $O(|V|^3)$. In order to examine the time complexity of such indices which are based on determining shortest paths for every vertex in a graph, e.g., I_{loc}^j, one can argue almost analogously. Further, it can be similarly shown that the remaining information measures possess polynomial time complexity. The computational complexity of OdC and Cr (see next section) has already been discussed in [51,52].

Additional Network Complexity Measures

As stated in the introduction, we will additionally evaluate the uniqueness of the Offdiagonal complexity and the graph index Cr, see [51,52].

Definition 8. *Let $G=(V,E)$ be a graph and let $(c_{ij})_{ij}$ be the vertex-vertex link correlation matrix, see [52]. c_{ij} denotes the number of all neighbors with degree $j>i$ of all vertices with degree i [51]. δ_{max} stands for the maximum degree of G. The normalized version of OdC can be defined as [51]*

$$OdC := \frac{-\left(\sum_{|V|=0}^{\delta_{max}-1} \tilde{a}_{|V|} \log\left(\tilde{a}_{|V|}\right)\right)}{\log(|V|-1)} \in [0,1], \tag{26}$$

where

$$\tilde{a}_{|V|} := \frac{a_{|V|}}{\sum_{|V|=0}^{\delta_{max}-1} a_{|V|}}, \tag{27}$$

and

$$a_{|V|} := \sum_{i=1}^{\delta_{max}-|V|} c_{i,i+|V|}, \mathit{see}\ [51]. \tag{28}$$

Definition 9. *Let $G=(V,E)$ be a graph and let r be the largest eigenvalue computed from its adjacency matrix.*

$$Cr := 4c_r(1-c_r) \in [0,1], \tag{29}$$

where

$$c_r := \frac{r - 2\cos\left(\frac{\pi}{|V|+1}\right)}{|V| - 1 - 2\cos\left(\frac{\pi}{|V|+1}\right)}. \tag{30}$$

Before discussing numerical results, we describe the databases and our developed software in brief.

Chemical Graph Databases

- MS 2265: This database has been extracted by own software from the commercially available mass spectral database NIST [75]. It contains 2265 selected chemical structures with different skeletons originating from the database NIST. This database has been already used in [63] for investigating different aspects of topological descriptors. It holds $4 \leq |V| \leq 19$; $2 \leq \rho(G) \leq 15 \; \forall \; G \in$ MS 2265.
- AG 3982: The original freely available database called Ames Genetoxicity contains 6512 chemical compounds, see [76,77]. After filtering the isomorphic graphs by using SubMat [78], we obtained 3982 structurally different skeletons, that is, all atoms and all bonds are considered as equal. The database was created from six different public sources [76,77]. Each structure has a class label (0 and 1) that results from the so-called Ames test indicating the genetoxicity of a substance. So far, the mentioned test has often been used in pharmaceutical sciences when investigating new molecules [76]. It holds $2 \leq |V| \leq 109$; $1 \leq \rho(G) \leq 47 \; \forall \; G \in$ AG 3982.
- APL 91075: The ASINEX Platinum Collection is a freely available, in-house designed and synthesized collection of 126615 drug-like compounds [79,80]. The filtering process of the isomorphic graphs by using a Python program resulted in 91075 structurally different skeletons. A notable feature of this database is that it contains structures from chemical subareas which are often under-represented in other available structure libraries [80]. Here, the chemical structures represent unlabeled and undirected graphs (skeletons). It holds $6 \leq |V| \leq 60$; $3 \leq \rho(G) \leq 36 \; \forall \; G \in$ APL 91075.
- C15 trees: This synthetic graph class [63] consists of 4347 alkane isomers with 15 carbon atoms (vertices). By definition, trees are connected, cycle free and here represent unlabeled and undirected graphs (skeletons). This database has been created by the software Molgen, see also [63].
- C15 ring 1: This synthetic graph class [63] consists of 60077 hydrocarbon isomers with 15 carbon atoms (vertices) containing one ring (cycle) and only single bonds. Hence, the structures can be treated as unlabeled and undirected graphs (skeletons). This database has been created by the software Molgen, see also [63].
- C15 ring 2: This synthetic graph class [63] consists of 94013 hydrocarbon isomers with 15 carbon atoms (vertices) containing two rings (cycles) and only single bonds. Hence, the structures can be treated as unlabeled and undirected graphs (skeletons). This database has been created by the software Molgen, see also [63].

Software and Data Processing

In order to generate and process our chemical graphs, we used the known Molfile format [81]. The database AG 3982 was originally available in Smiles format that we converted to Molfile format (SDF) using a Python procedure. The databases MS 2265 and APL 91075 were directly available in Molfile format (SDF). To apply the information-theoretic measures to the previously presented graph databases, we performed a procedure to filter all isomorphic graphs contained in these databases. This isomorphism check was done by applying the software SubMat [78] and the previously mentioned Python program. As a result, we obtained sets of graphs containing different skeletons representing the underlying graph topology of the molecules.

We implemented all used topological measures in Python using freely available libraries like Networkx, Openbabel and Pybel packages [82]. For the calculations we have performed in this paper, we started from the Molfile representation of a chemical structure, created the corresponding adjacency matrix and computed the topological indices based on the developed Python program. The databases containing the synthetic graph structures (isomers) have been generated by the software Molgen, see also [63].

Results and Discussion

In this section, we will apply the complexity measures presented in the previous section. As stated before, we mainly put the emphasis on exploring the relatedness between the topological information content I_{orb} and our graph entropy measures $I^{\lambda}_{f^V_{lin}}$ and $I^{\lambda}_{f^V_{exp}}$. Moreover, we numerically calculate further information-theoretic network measures presented in the last section and interpret the results. In particular, an interesting question will be to investigate the so-called uniqueness of the measures when applying them to both databases containing real and synthetic chemical graphs.

Numerical Results

In the following, we discuss and interpret numerical results when applying the selected descriptors to sets containing real chemical structures. Our study involves calculating and interpreting dependency plots, cumulative entropy distributions, and the so-called uniqueness of the used topological indices [66].

Relatedness between I_{orb} and $I^{\lambda}_{f^{V}_{lin}}, I^{\lambda}_{f^{V}_{exp}}$. We start to examine how the entropies I_{orb} and $I^{\lambda}_{f^{V}_{lin}}, I^{\lambda}_{f^{V}_{exp}}$ capture structural information of our graphs and depict the scatter plots (see Figure (1) and Figure (2)) for exploring the correlation between the measures. To tackle this problem, we now only consider Figure (1) exemplarily. Clearly, the main observation is that I_{orb} is highly uncorrelated with $I^{\lambda}_{f^{V}_{lin}}, I^{\lambda}_{f^{V}_{exp}}$. In order to interpret this figure in more detail, we select the graphs marked by red-colored arrows (these graphs are depicted in Figure (3), (4), (5)) whose entropies (for practical scaling reasons, we always calculated normalized entropies) are extremal with respect to I_{orb} or $I^{\lambda}_{f^{V}_{lin}}, I^{\lambda}_{f^{V}_{exp}}$. Before discussing the results, we give two mathematical statements [27,64].

Proposition 1. *If G is vertex transitive [27,57], then $I_{orb}(G)=0$*

Proposition 2. *If G is k-regular [57], then $I_{f^V}(G)=\log(|V|)$ and, hence, $I^{\lambda}_{f^V}(G)=0$.*

The graph $G=C_6$ with $I_{orb}(C_6)=0$ and $I^{\lambda}_{f^{V}_{exp}}(C_6)=0$ is a cycle possessing six vertices (Figure (1)). Because C_6 is vertex transitive, there is only one orbit containing all vertices and, thus, according to Proposition (1), we get $I_{orb}(C_6)=0$. Moreover, C_6 is 2-regular. Applying Proposition (2) yields to $I_{f^{V}_{exp}}(C_6)=\log(6)$ (see also Equation (22)) and, hence, $I^{\lambda}_{f^{V}_{exp}}(C_6)=\lambda(\log(6)-\log(6))=0$.

The interrelation between the entropies (I_{orb} and $I^{\lambda}_{f^{V}_{exp}}$) for the graph depicted by Figure (3) can be understood by applying the previously stated propositions. As we easily see, this graph has a cyclic and symmetric structure and, therefore, I_{orb} is low. For the same reason when explaining the interrelation for the fully cyclic graph C_6, the corresponding entropy value of $I^{\lambda}_{f^{V}_{exp}}$ is also low. The next entropy relation we want to describe concerns the graph G (see Figure (4)) whose topological information content is relatively high and $I^{\lambda}_{f^{V}_{exp}}(G)=1$. Here, $I^{\lambda}_{f^{V}_{exp}}(G)=1$ means that the entropy $I_{f^{V}_{exp}}(G)$ attains a minimum. The reason why the topological information content is relatively high for this graph can be understood by the fact that the degree of symmetry is rather low resulting in the observation that most of the vertex orbits of G are only singleton partitions. The last graph G we will inspect possesses $I_{orb}(G)=1$ and a relatively small value of $I^{\lambda}_{f^{V}_{lin}}$. This graph G (see Figure (5)) is an element of a certain subset that is highlighted by the red-colored rectangle in Figure (1). To determine I_{orb} for this graph, we have to calculate the partitions according to the equivalence criterion that is based on vertex orbits. At first glance, G seems to be symmetric (according to this criterion) but a deeper inspection leads to the result that all vertex orbits are singleton partitions. Hence, $I_{orb}(G)=1$. But based on the cyclic structure of G and again by definition of $I^{\lambda}_{f^{V}_{lin}}$ and Proposition (2), we infer that its corresponding entropy value is relatively small.

Uniqueness of the Descriptors. Besides investigating the problem how the measures capture structural information of the considered chemical structures, we now examine another important property of a topological index, namely the ability to discriminate the graphs as unique as possible. This characteristic property of a structural graph measure is often referred to as degeneracy [66,69,83]; related work can be found in, e.g., [63,66,69,83,84]. To evaluate the uniqueness of a measure I, we

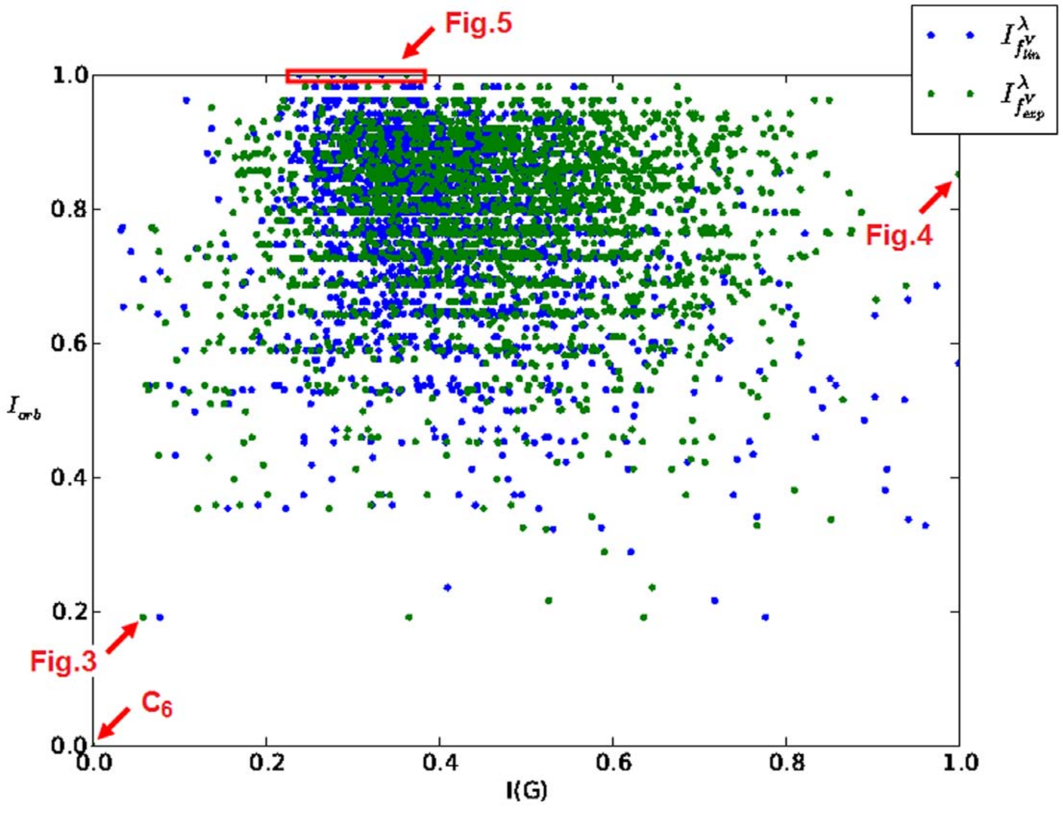

Figure 1. I_{orb} versus $I^{\lambda}_{f^{V}_{lin}}, I^{\lambda}_{f^{V}_{exp}}$ for MS 2265. (reference label: scatter_plot1).
doi:10.1371/journal.pone.0008057.g001

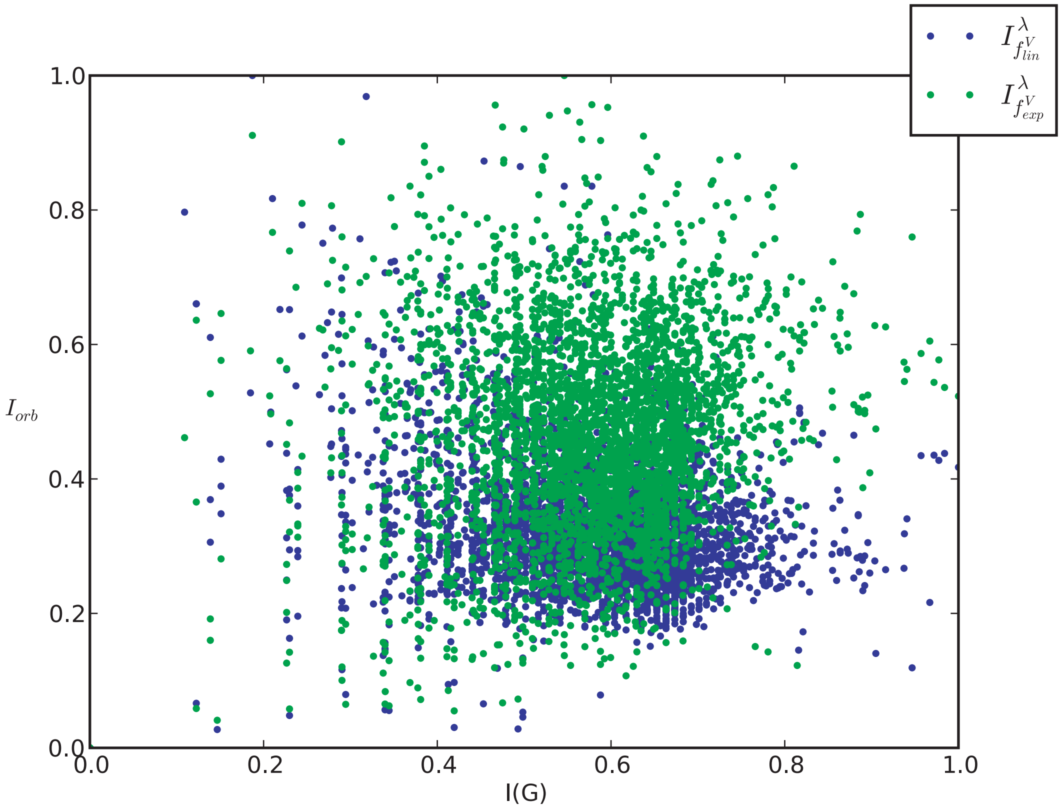

Figure 2. I_{orb} versus $I^{\lambda}_{f^V_{lin}}, I^{\lambda}_{f^V_{exp}}$ for AG 3982. (reference label: scatter_plot2).
doi:10.1371/journal.pone.0008057.g002

here apply the sensitivity index proposed by Konstantinova [66]:

$$S(I) = \frac{|\mathcal{G}| - |\mathcal{G}_j|}{|\mathcal{G}|}. \tag{31}$$

I denotes a topological index and $\mathcal{G}$ denotes a set of arbitrary graphs, respectively. $|\mathcal{G}_j|$ stands for the number of graphs $G_i \in \mathcal{G}$ which can not be distinguished by calculating I. If it holds $S(I) = 1$, we know by definition that it does not exist any pair of non-isomorphic graphs $G_i \in \mathcal{G}$ possessing the same value of I.

We now start discussing the results shown in Table (1) when evaluating the sensitivity of our indices and start with the topological information content I_{orb}. We note that the sensitivity values depends on the chosen decimal places. Here, we calculated $S(I)$ with an

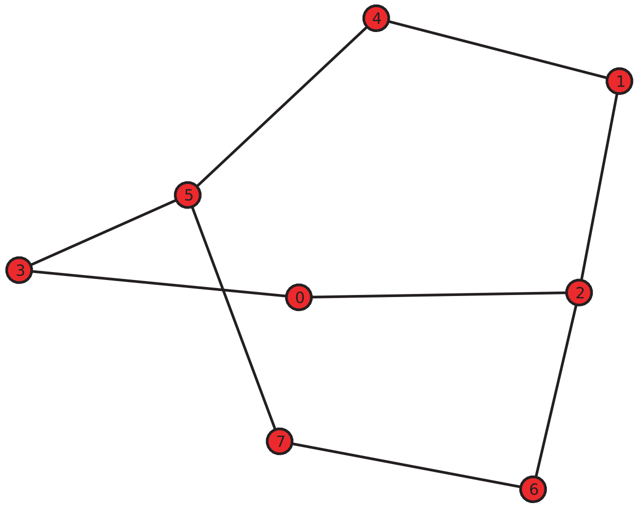

Figure 3. Example Graph with relatively small value of both I_{orb} and $I^{\lambda}_{f^V_{exp}}$. (reference label: graph_plot1).
doi:10.1371/journal.pone.0008057.g003

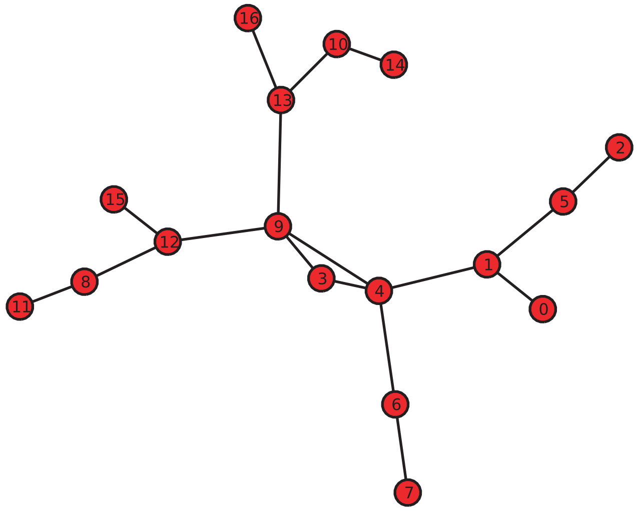

Figure 4. Example Graph G with relatively large value of I_{orb} and $I^{\lambda}_{f^V_{exp}}(G) = 1$. (reference label: graph_plot2).
doi:10.1371/journal.pone.0008057.g004

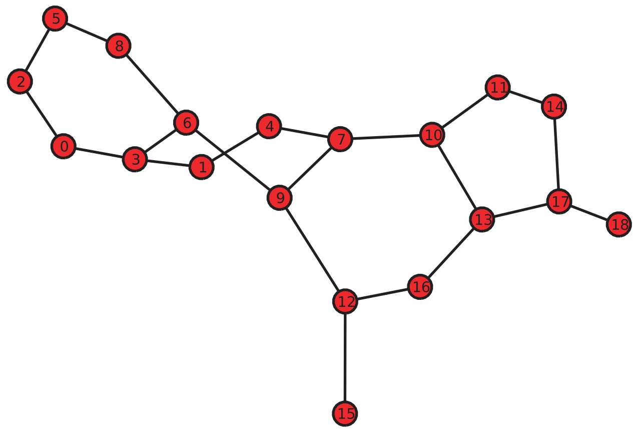

Figure 5. Example Graph G with $I_{orb}(G)=1$ and relatively small $I^{\lambda}_{f^V_{exp}}$. (reference label: graph_plot3).
doi:10.1371/journal.pone.0008057.g005

accuracy of 6 decimal places. First, we see that I_{orb} has a very low discrimination power compared to the remaining information measures, except OdC and Wiener index. This can be understood by briefly recalling the definition of the topological information content (see also Remark(1)): The main idea is to partition the vertex set in equivalence classes according to the criterion that each such class contains topologically equivalent vertices [2,19]. Therefore, this measure is based on symmetry with respect to the topologically equivalent vertices having the same degrees (the vertices to be in the same vertex orbit must have the same degree). Thus, we can easily construct graphs having the same vertex orbits but whose underlying topology is different, and, evidently, the uniqueness of I_{orb} is often very low. Interestingly, OdC has similarly to I_{orb} a very low discrimination power. This can be explained by arguing that the underlying basis for calculating this measure - the vertex-vertex link correlation matrix - does not capture complex structural features of a graph adequately (at least for the considered graph classes). As known and reflected by Table (1), the uniqueness of the Wiener index is also very low [66]. In contrast to this, the sensitivity values of Cr for MS 2265 and AG 3982 are feasible. But for APL 91075, its uniqueness is very low. This clearly shows that the uniqueness of a topological index strongly depends on the graph class (structural diversity of graphs) under consideration (see also "Summary and Conclusion" section). Note that the sensitivity calculation of our information indices $I^{\lambda}_{f^V_{lin}}, I^{\lambda}_{f^V_{exp}}$ led to much better results. By choosing the coefficients exponentially decreasing (see Equation (21)), the resulting entropy measure is able to discriminate all graphs of MS 2265 uniquely and, hence, $S(I^{\lambda}_{f^V_{exp}})=1$. For AG 3982 and APL 91075, we obtained that 12 and 220 graphs could not be distinguished when applying $I^{\lambda}_{f^V_{exp}}$, respectively. The sensitivity evaluation of $I^{\lambda}_{f^V_{lin}}$ led to quite similar result. In summary, Table (1) shows that our information indices possess a very high uniqueness for all three chemical databases and, therefore, can discriminate real chemical graphs successfully. A more mathematical explanation for this result is as follows: Instead of determining partitions by using a graph invariant, e.g., number of vertices or edges, and then calculating a probability for each such partition, we assign a probability value to every vertex in a graph. By using our proposed information functional, we furthermore compute the full topological neighborhood of all involved vertices (atoms) of the structure [63]. To determine the entropy of the underlying graph topology, the vertex probabilities (see Equation (24)) can be interpreted as percentage rates of the entire graph structure for every vertex instead of lumping structural properties together when calculating the partitions (according to a an equivalence criterion). As a conclusive remark, we want to emphasize that I_U and some other computed information indices also possess a high discrimination power (see Table (1)).

To interpret the sensitivity values when applying our information measures to synthetical chemical graphs, we look at Table (2). Here, we applied the same graph measures to the presented synthetic graph classes. As before, the uniqueness of OdC, I_{orb} and W is for all three graph classes extremely low. Compared to W, one sees that Cr has a much better discrimination power. By exemplarily determining the number of graphs which could not be distinguished by Cr for C15 ring 2, we yield $|\mathcal{G}_j|=47772$ (see Equation (31)).

However for the tree class, our $I^{\lambda}_{f^V_{exp}}$ discriminates all 4347 trees uniquely. Moreover, one observes that the sensitivity values of the remaining information measures for this graph class are high. The

Table 1. Calculation of sensitivity index $S(I)$ for chemical databases.

Topological index I	$S(I)$ for MS 2265	$S(I)$ for AG 3982	$S(I)$ for APL 91075
OdC	0.142604	0.247363	0.029744
$I^{\lambda}_{f^V_{lin}}$	0.997350	0.995981	0.988723
$I^{\lambda}_{f^V_{exp}}$	1.0	0.996986	0.997584
I_{orb}	0.026931	0.074334	0.002723
I_D	0.859602	0.938724	0.873873
I_D^W	0.883885	0.947513	0.933033
I_U	0.999116	0.999497	0.996618
I_W	0.990286	0.990959	0.522799
I^1_{loc}	0.995584	0.994977	0.914389
I^2_{loc}	0.999116	0.996986	0.916453
I^3_{loc}	0.989403	0.973882	0.595783
W	0.014128	0.037920	0.001065
Cr	0.864017	0.919638	0.223892

doi:10.1371/journal.pone.0008057.t001

Table 2. Calculation of sensitivity index $S(I)$ for synthetic graph classes.

Topological index I	$S(I)$ for C15 trees	$S(I)$ for C15 ring 1	$S(I)$ for C15 ring 2
OdC	0.001380	0.000065	0.000031
$I^{\lambda}_{f^V_{lin}}$	0.983897	0.963713	0.980034
$I^{\lambda}_{f^V_{exp}}$	1.0	0.998601	0.997383
I_{orb}	0.001380	0.000116	0.000042
I_D	0.634000	0.124972	0.093774
I^W_D	0.748562	0.142567	0.108889
I_U	0.998159	0.937213	0.859530
I_W	0.987577	0.771842	0.586365
I^1_{loc}	0.965263	0.568736	0.394817
I^2_{loc}	0.965033	0.669940	0.553370
I^3_{loc}	0.982286	0.785658	0.727622
W	0.000920	0.000116	0.000085
Cr	0.459627	0.502771	0.491857

doi:10.1371/journal.pone.0008057.t002

final result we want to emphasize is that by applying our information-based topological descriptors $I^{\lambda}_{f^V_{lin}}, I^{\lambda}_{f^V_{exp}}$, we obtained constantly high sensitivity values for all three synthetic graphs classes. In order to calculate the number of graphs which could not be distinguished by $I^{\lambda}_{f^V_{exp}}$ and I_U, we choose again the class C15 ring 2. For $I^{\lambda}_{f^V_{exp}}$, we get $|\mathcal{G}_j|=246$ but by applying I_U, we yield $|\mathcal{G}_j|=13206$.

Cumulative Entropy Distributions. The cumulative entropy distributions are illustrated by Figure (6). In these plots, the x-axis represents the normalized entropy values whereas the y-axis shows the percentage rate of chemical graphs having a (normalized) entropy value less or equal $I(G)$. We want to remark that the measures were normalized by using $\tilde{I}=\frac{I-\min(I)}{\max(I)-\min(I)}$.

We start by observing that about 80% of the graphs of MS 2265 possess relatively small entropy values when evaluating I_W (see Equation (9)). In contrast, 80% of the graphs have large entropy values by calculating $I^W_D, I^1_{loc}, I^2_{loc}I_{orb}$ (see Equation (6), (17), (4)). This result can be interpreted such that the measures capture structural information of the graphs quite differently because the corresponding entropy distributions are almost reverse. The

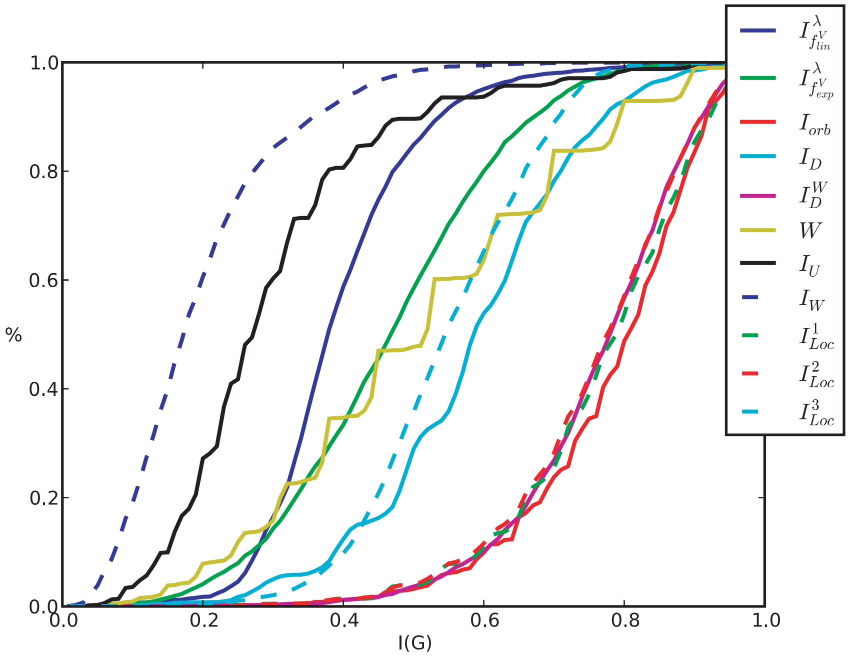

Figure 6. Cumulative Entropy Distributions for MS 2265. (reference label: cum_plot1).
doi:10.1371/journal.pone.0008057.g006

interrelation between the graph entropies $I^{\lambda}_{f^V_{lin}}, I^{\lambda}_{f^V_{exp}}$ (see Equation (23)) and I_{orb} is quite similar to the just described one. Finally, note that the findings of the section where we have examined the relatedness between the selected measures support this hypothesis.

Equally, the cumulative entropy distributions of AG 3982 are depicted in Figure (7). One can see that for some indices the curve progressions appear quite diversely, e.g., W, I_U, I_W, I^1_{loc}. A possible explanation for this could be the fact that AG 3982 is structurally more diverse than MS 2265. For the remaining entropy measures, the situation is similar as described in Figure (6). Interestingly, the cumulative similarity distribution of the discussed information measures illustrated by Figure (6) and Figure (8) are again quite similar.

In particular, we have found that for all three chemical databases, the evaluation of the topological information content (see Equation (4)) and the partition-independent measures (see Equation (23)) led to clearly different cumulative entropy distributions that is obviously in accordance with the results of the preceding sections.

Summary and Conclusion

In the present paper, we studied interrelations between classical and novel entropy measures to quantify the structural information content of networks. Here, these measures served as graph complexity measures which take certain structural features of the networks under consideration into account. In the following, we express the main findings of the paper in brief:

- We explored the relatedness between information measures for graphs. In particular, we examined the correlation between the topological information content I_{orb} (see Equation (4)) and the partition-independent measures $I^{\lambda}_{f^V}(G)$ (see Equation (23)) by interpreting the corresponding scatter plots. Let G be a graph. If $I_{orb}(G)$ is small or even zero, then G is symmetric with respect to topologically equivalent vertices having the same degrees which form the so-called vertex orbits. Then, if the value of $I^{\lambda}_{f^V}(G)$ is also small, G has a cyclic structure and represents a graphs that is equal or very similar to a k-regular graph. As shown in Figure (5), a graph G whose value of $I_{orb}(G)$ is large can be also cyclic and, hence, possesses a small $I^{\lambda}_{f^V}(G)$ value. Further, for a graph G whose value of $I^{\lambda}_{f^V}(G)$ is large (Figure (4)), the involved mean information content $I_{f^V}(G)$ is low or even attains a minimum. In [63], we showed that such graphs typically represent chain-like graphs or generally speaking, graphs with a low branching factor. The reason why $I^{\lambda}_{f^V}(G)$ has small values for graphs containing cyclic structures seems (which are symmetric) logical because it corresponds to the accepted concept [3] that symmetry leads to a decrease of complexity.
- Another important aspect of our numerical study was to examine the discrimination power of the used network measures. We found that the topological information content I_{orb} was weak in distinguishing non-isomorphic graphs, i,e., it's sensitivity value was very low. In contrast, the sensitivity evaluation for our partition-independent measures $I^{\lambda}_{f^V}(G)$ led to constantly good results when applying the measures to real and synthetic chemical structures. Recall that a high uniqueness of a complexity measure corresponds to the ability to distinguish networks whose structural similarity is very high. Hence, this feature could be useful (as future work) when considering graphs which were inferred statistically (erroneous graphs) [85]. As an important remark, we want to emphasize that the uniqueness of a topological index also depends on the considered graphs class. Note that our chemical graphs are particularly small and structurally not very diverse compared to the ones used in e.g., [60]. Especially for those graphs whose numbers of vertices are rather small, highly discriminative measures are extremely important for quantifying structural information as unique as possible. That is one reason why we studied the uniqueness of topological indices for chemical graph analysis. A further reason relates to the fact that descriptors with a high discrimination power are often useful for QSPR/QSAR. But we have already seen that an index I does not necessarily perform well for several graph classes at

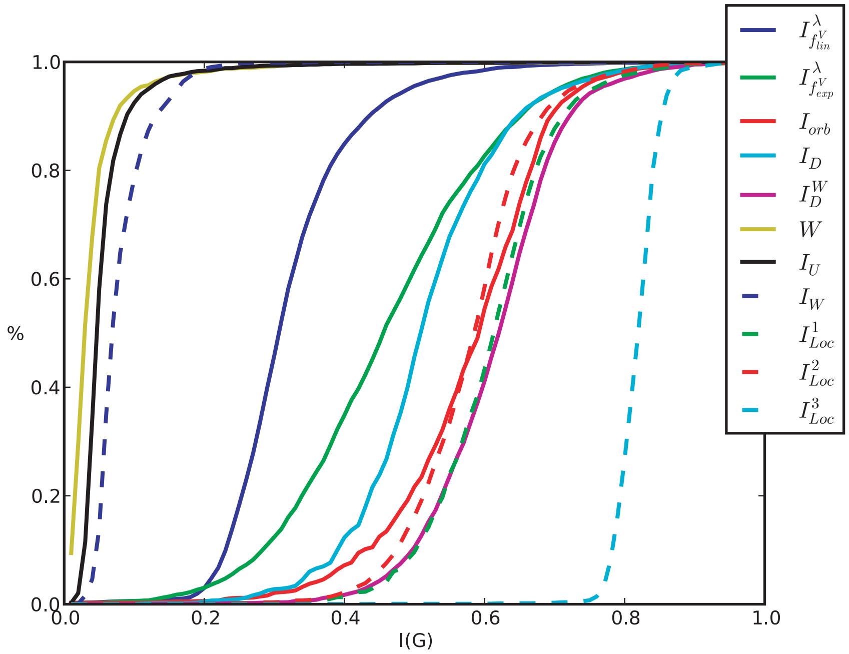

Figure 7. Cumulative Entropy Distributions for AG 3982. (reference label: cum_plot2).
doi:10.1371/journal.pone.0008057.g007

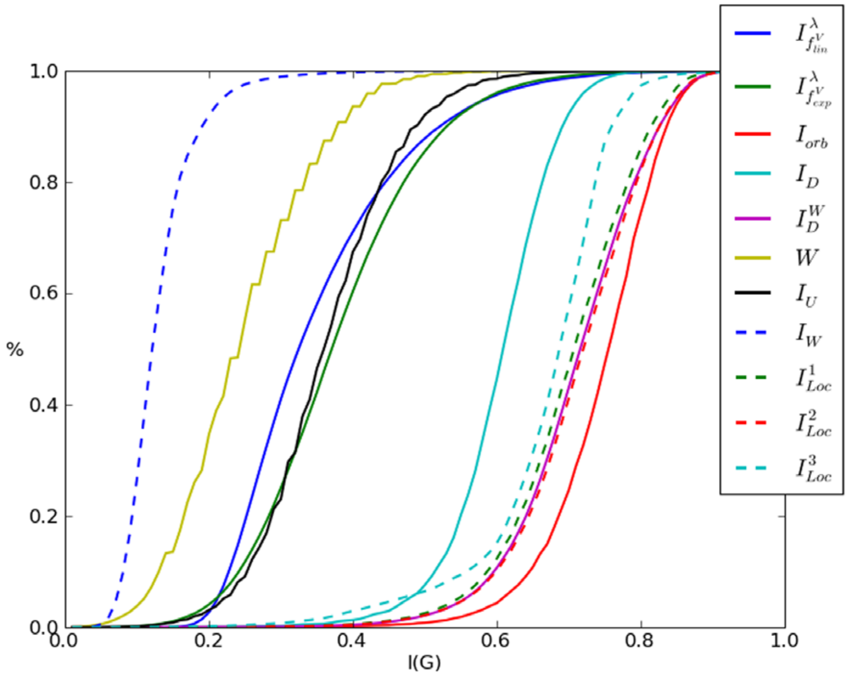

Figure 8. Cumulative Entropy Distribution for APL 91075. (reference label: cum_plot3).
doi:10.1371/journal.pone.0008057.g008

the same time. To further shed light on this problem, we briefly pick up the first argument of this paragraph. In this paper and in [84], we evaluated the uniqueness of some information-theoretic measures for real and synthetic chemical structures. For some indices, e.g., I_D^W, I_D which performed very well for real chemical graphs, we got worse results when applying these measures to synthetic graphs, e.g., isomers having 10 [84] and 15 vertices each.

- For the real chemical databases, the cumulative entropy distributions of some measures were calculated. This approach can be considered as an important preprocessing step to learn how the measures capture structural information of networks. Particularly, it is suitable to explore certain correlations between the measures and, finally, to learn whether the complexity indices capture structural information differently or similarly.

As a conclusive remark, we emphasize that the presented information-theoretic methods to analyze complex networks bear a considerable potential. Our study aimed to get a better understanding towards the problem of characterizing chemical graphs using information-theoretic complexity measures. In this paper, we put the emphasis on such measures which have already been applied in the context of mathematical chemistry and drug design. We think that our results can help to apply the measures to more complex network classes and to interpret the results more adequately than before.

In the future, we want to extend our measures for determining the structural complexity of weighted chemical graphs (i.e., incorporating atom and bond types) and test their ability to tackle QSAR/QSPR problems. Further, we would like to test novel information indices by combining existing ones and evaluate their discrimination power. Moreover, an interesting task would be to classify molecules by using this approach and to apply it to special problems in drug design.

Acknowledgments

We thank Katja Hansen for providing the Ames database and Danail Bonchev, Frank Emmert-Streib, Alexander Mehler, and Abbe Mowshowitz for fruitful discussions.

Author Contributions

Conceived and designed the experiments: MD NB KV AG. Performed the experiments: MD NB KV AG. Analyzed the data: MD NB KV AG. Contributed reagents/materials/analysis tools: MD NB KV AG. Wrote the paper: MD NB KV AG.

References

1. Basak SC (1999) Information-theoretic indices of neighborhood complexity and their applications. In: Balaban AT, Devillers J, eds. Topological Indices and Related Descriptors in QSAR and QSPAR, Gordon and Breach Science Publishers. pp 563–595. Amsterdam, The Netherlands.
2. Bonchev D (1983) Information Theoretic Indices for Characterization of Chemical Structures. Chichester: Research Studies Press.
3. Bonchev D (2003) Complexity in Chemistry. Introduction and Fundamentals. Taylor and Francis. Boca Raton, FL, USA.
4. Sommerfeld E, Sobik F (1994) Operations on cognitive structures - their modeling on the basis of graph theory. In: Albert D, ed. Knowledge Structures, Springer. pp 146–190.
5. Mehler A (2009) A quantitative graph model of social ontologies by example of wikipedia. In: Mehler A, Sharoff S, Rehm G, Santini M, eds. Genres on the Web: Computational Models and Empirical Studies, Springer. To appear.
6. Minoli D (1975) Combinatorial graph complexity. Atti Accad Naz Lincei, VIII Ser, Rend, Cl Sci Fis Mat Nat 59: 651–661.
7. Ulanowicz RE (2001) Information theory in ecology. Computers and Chemistry 25: 393–399.
8. Bonchev D, Rouvray DH (2005) Complexity in Chemistry, Biology, and Ecology. Mathematical and Computational Chemistry. New York, NY, USA: Springer.

9. Dancoff SM, Quastler H (1953) Information content and error rate of living things. In: Quastler H, ed. Essays on the Use of Information Theory in Biology, University of Illinois Press. pp 263–274.
10. Linshitz H (1953) The information content of a battery cell. In: Quastler H, ed. Essays on the Use of Information Theory in Biology, University of Illinois Press. Urbana, IL, USA.
11. Morowitz H (1953) Some order-disorder considerations in living systems. Bull Math Biophys 17: 81–86.
12. Trucco E (1956) A note on the information content of graphs. Bull Math Biol 18: 129–135.
13. Shannon CE, Weaver W (1997) The Mathematical Theory of Communication. Urbana, IL, USA: University of Illinois Press.
14. Emmert-Streib F, Dehmer M (2007) Information theoretic measures of UHG graphs with low computational complexity. Appl Math Comput 190: 1783–1794.
15. 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.
16. Mowshowitz A (1968) Entropy and the complexity of graphs II: The information content of digraphs and infinite graphs. Bull Math Biophys 30: 225–240.
17. Mowshowitz A (1968) Entropy and the complexity of graphs III: Graphs with prescribed information content. Bull Math Biophys 30: 387–414.
18. Mowshowitz A (1968) Entropy and the complexity of graphs IV: Entropy measures and graphical structure. Bull Math Biophys 30: 533–546.
19. Rashevsky N (1955) Life, information theory, and topology. Bull Math Biophys 17: 229–235.
20. Basak SC, Magnuson VR (1983) Molecular topology and narcosis. Arzeim-Forsch/Drug Design 33: 501–503.
21. Bonchev D (1979) Information indices for atoms and molecules. Commun Math Comp Chem 7: 65–113.
22. Bonchev D, Mekenyan O, Trinajstić N (1981) Isomer discrimination by topological information approach. J Comp Chem 2: 127–148.
23. Devillers J, Balaban AT (1999) Topological Indices and Related Descriptors in QSAR and QSPR. Amsterdam, The Netherlands: Gordon and Breach Science Publishers.
24. Diudea MV, Gutman I, Jäntschi L (2001) Molecular Topology. Ney York, NY, USA: Nova Publishing.
25. Todeschini R, Consonni V, Mannhold R (2002) Handbook of Molecular Descriptors. Weinheim, Germany: Wiley-VCH.
26. Konstantinova EV, Skorobogatov VA, Vidyuk MV (2002) Applications of information theory in chemical graph theory. Indian Journal of Chemistry 42: 1227–1240.
27. Mowshowitz A, Mitsou V (2009) Entropy, orbits and spectra of graphs. In: Dehmer M, Emmert-Streib F, eds. Analysis of Complex Networks: From Biology to Linguistics, Wiley-VCH. pp 1–22.
28. 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.
29. Constantine G (1990) Graph complexity and the laplacian matrix in blocked experiments. Linear and Multilinear Algebra 28: 49–56.
30. Jukna S (2006) On graph complexity. Comb Probab Comput 15: 855–876.
31. Kolmogorov AN (1965) Three approaches to the definition of information. Probl Peredaci Inform 1: 3–11.
32. Li M, Vitányi P (1997) An Introduction to Kolmogorov Complexity and Its Applications Springer.
33. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks. From Biological Networks to the Internet and WWW Oxford University Press.
34. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
35. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509–512.
36. Albert R, Barabási AL, Jeong H, Bianconi G (2000) Power-law distribution of the world wide web. Science 287: 130–131.
37. Erdös P, Rényi P (1960) On the evolution of random graphs. Magyar Tud Akad Mat Kutató Int Közl 5: 17–61.
38. Bornholdt S, Schuster HG (2003) Handbook of Graphs and Networks: From the Genome to the Internet. New York, NY, USA: John Wiley & Sons, Inc.
39. Adamic L, Huberman B (2000) Power-law distribution of the world wide web. Science 287: 2115a.
40. Mason O, Verwoerd M (2007) Graph theory and networks in biology. IET Systems Biology 1: 89–119.
41. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, et al. (2000) Graph structure in the web: Experiments and models. In: Proceedings of the 9-th WWW Conference. Amsterdam.
42. Skorobogatov VA, Dobrynin AA (1988) Metrical analysis of graphs. Commun Math Comp Chem 23: 105–155.
43. Watts DJ (1999) Small worlds: The dynamics of networks between order and randomness. Princeton, NJ, USA: Princeton University Press.
44. da F Costa L, Rodrigues F, Travieso G (2007) Characterization of complex networks: A survey of measurements. Advances in Physics 56: 167–242.
45. Barabási AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. Nature Reviews Genetics 5: 101–113.
46. Brandes U, Erlebach T (2005) Network Analysis. Lecture Notes in Computer Science. Berlin Heidelberg New York: Springer.
47. Mehler A (2006) In search of a bridge between network analysis in computational linguistics and computational biology – A conceptual note. In: Proceedings of the 2006 International Conference on Bioinformatics & Computational Biology (BIOCOMP'06), 2006, Las Vegas, Nevada, USA. pp 496–500.
48. Newman MEJ (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103: 8577–8582.
49. Koschützki D, Lehmann KA, Peters L, Richter S, Tenfelde-Podehl D, et al. (2005) Clustering. In: Brandes U, Erlebach T, eds. Centrality Indices, Springer, Lecture Notes of Computer Science. pp 16–61.
50. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences Cambridge University Press.
51. Kim J, Wilhelm T (2008) What is a complex graph? Physica A 387: 2637–2652.
52. Claussen JC (2007) Characterization of networks by the offdiagonal complexity. Physica A 365–373: 321–354.
53. Wilhelm T, Brueggemann R (2001) Information theoretic measures for the maturity of ecosystems. In: Matthies M, Malchow H, Kriz J, eds. Integrative Systems Approaches to Natural and Social Sciences - Systems Science 2000, Springer. pp 263–273. Berlin, Germany.
54. Wilhelm T, Hollunder J (2007) Information theoretic description of networks. Physica A 388: 385–396.
55. Solé RV, Valverde S (2004) Information theory of complex networks: On evolution and architectural constraints. In: Lecture Notes in Physics, volume 650. pp 189–207.
56. Antiqueira L, da F Costa L (2009) Characterization of subgraph relationships and distribution in complex networks. New Journal of Physics 11.
57. Harary F (1969) Graph Theory. Reading, MA, USA: Addison Wesley Publishing Company.
58. Garey MR, Johnson DS (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness. Series of Books in the Mathematical Sciences W. H. Freeman.
59. da F Costa L, Rodrigues FA, Hilgetag CC, Kaiser M (2009) Beyond the average: Detecting global singular nodes from local features in complex networks. Europhysics Letters 87: 18008(1)–18008(6).
60. da F Costa L, Rodrigues FA (2009) Seeking for simplicity in complex networks. Europhysics Letters 85: 48001(1)–48001(6).
61. Bonchev D (2000) Overall connectivities and topological complexities: A new powerful tool for QSPR/QSAR. J Chem Inf Comput Sci 40: 934–941.
62. Randić M, Plavšić DP (2002) On the concept of molecular complexity. Croatica Chemica Acta 75: 107–116.
63. 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 Inf Model 49: 1655–1663.
64. Dehmer M (2008) A novel method for measuring the structural information content of networks. Cybernetics and Systems 39: 825–843.
65. Dehmer M, Emmert-Streib F (2008) Structural information content of networks: Graph entropy based on local vertex functionals. Comput Biol Chem 32: 131–138.
66. Konstantinova EV (2006) On some applications of information indices in chemical graph theory. In: Ahlswede R, Bäumer L, Cai N, Aydinian H, Blinovsky V, et al., eds. General Theory of Information Transfer and Combinatorics, Springer, Lecture Notes of Computer Science. pp 831–852.
67. Emmert-Streib F, Dehmer M (2009) Information processing in the transcriptional regulatory network of yeast: Functional robustness. BMC Syst Biol 3.
68. Hirata H, Ulanowicz RE (1984) Information theoretical analysis of ecological networks. Int J Syst Sci 15: 261–270.
69. Bonchev D, Trinajstić N (1977) Information theory, distance matrix and molecular branching. J Chem Phys 67: 4517–4533.
70. Brillouin L (1956) Science and Information Theory. New York: Academic Press.
71. McKay BD (1981) Graph isomorphisms. Congressus Numerantium 730: 45–87.
72. Trinajstić N (1992) Chemical Graph Theory. Boca Raton, FL, USA: CRC Press.
73. Dehmer M, Emmert-Streib F (2009) Towards network complexity. In: Zhou J, ed. Complex Sciences, Springer, Berlin/Heidelberg, Germany, volume 4 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. pp 707–714.
74. Chepoi V, Klavzar S (1997) The wiener index and the szeged index of benzenoid systems in linear time. Journal of Chemical Information and Computer Sciences 37: 752–755.
75. Stein SE (1998) NIST, Mass spectral database 98. www.nist.gov/srd/nist1a.htm. National Institute of Standards and Technology, Gaithersburg, MD, USA.
76. Schwaighofer A, Schroeter T, Mika S, Hansen K, Laak AT, et al. (2008) A probabilistic approach to classifying metabolic stability. J Chem Inf Model 48: 785–796.
77. Hansen K, Mika S, Schroeter T, Sutter A, Laak AT, et al. (2009) A benchmark data set for in silico prediction of ames mutagenicity. J Chem Inf Model.
78. Scsibrany H, Varmuza K (2004) Software SubMat. www.lcm.tuwien.ac.at. Vienna University of Technology, Institute of Chemical Engineering, Laboratory for Chemometrics, Austria.
79. Asinex (2008) ASINEX platinum collection. http://www.asinex.com.
80. Mukherjee P, Desai P, Ross L, White, Averya MA (2008) Structure-based virtual screening against sars-3clpro to identify novel non-peptidic hits. Bioorganic & Medicinal Chemistry 16: 4138–4149.
81. Gasteiger J, Engel T (2003) Chemoinformatics - A Textbook. Weinheim, Germany: Wiley VCH.

82. O'Boyle NM, Morley C, Hutchison GR (2008) Pybel: A python wrapper for the openbabel cheminformatics toolkit. Chemistry Central Journal 2.
83. Balaban AT, Ivanciuc O (1999) Historical development of topological indices. In: Balaban AT, Devillers J, eds. Topological Indices and Related Descriptors in QSAR and QSPAR, Gordon and Breach Science Publishers. pp 21–57. Amsterdam, The Netherlands.
84. Dehmer M, Varmuza K (2009) On aspects of the degeneracy of topological indices. submitted for publication.
85. Emmert-Streib F, Dehmer M (2008) Analysis of Microarray Data: A Network-Based Approach Wiley-VCH.