J Math Chem (2014) 52:805–819
DOI 10.1007/s10910-013-0294-9

ORIGINAL PAPER

Large-scale analysis of structural branching measures

Michael Schutte · Matthias Dehmer

Received: 23 October 2013 / Accepted: 11 November 2013 / Published online: 22 November 2013

Abstract Structural branching of graphs has been investigated extensively. Yet, no method/model has yet been developed which captures all aspects of branching meaningfully. Another shortcoming of nearly all related work in this area is the fact that only small sets of example graphs have been used to perform those studies. Instead, we investigate structural branching of graphs statistically by using large sets of exhaustively generated graphs. Our findings explain some of the limits of existing branching measures as well as the search for novel branching measures by using correlation analysis.

Keywords Molecular branching · Graphs · Graph analysis · Quantitative graph measures

1 Introduction

Examining complex structural features of graphs is a still ongoing problem in complex network analysis and has been investigated for several decades [10,13,28]. Nowadays, graph analysis can be divided into two major categories: Descriptive graph analysis [10,13,28] and Quantitative graph analysis [4,5,15,34]. The former relates to describe and characterize graphs by using structural properties. The latter deals with quantifying structural information of graphs by using certain measurements.

In this paper, we focus on aspects of quantitative graph analysis namely related to the structural interpretation of quantitative network measures [4,18]. Exploring this

M. Schutte · M. Dehmer (✉)
Institute for Bioinformatics and Translational Research, UMIT, Hall in Tirol, Austria
e-mail: matthias.dehmer@umit.at

M. Schutte
e-mail: michael.schutte@uiae.at

problem in general is still intricate as numerous graph measures have been developed whose properties are unexplored [18]. When focussing on information-theoretic graph measures, Dehmer [14] made already an attempt to classify the measures based on their structural interpretation. One important example thereof is *structural branching* of graphs which we are going to investigate in the present paper.

The main contribution of the paper is twofold: First, we perform a large-scale analysis by using exhaustively generated graphs to investigate structural branching. The idea to perform this study stems from the shortcomings of existing related work in this area: only small example data sets have been used and no rigorous statistical analysis of the problem has been yet performed. Second, we examine already existing measures towards their ability to detect branching meaningfully and discuss the limits of already existing branching measures/models by employing correlation analysis. Also, we find measures which turned out to be suitable branching measures but they were not identified as such before.

2 Related work

The problem of investigating branching of graphs has been explored graph-theoretically as well as by using so-called topological indices, see, e.g., [6,8,9,40]. An early attempt by using graph-theoretical and quantitative methods to quantify the degree of branching of graphs is due to Randić [40]. Randić focuses on trees consisting only of vertices of a degree not larger than 4, which correspond to the topology of alkane hydrocarbons. Randić notes that an intuitive treatment of branching is ambiguous even when the discussion is confined to relatively simple systems, see [40]. Then he proposed a topological descriptor χ, called branching index to formalize the concept [40]. χ has been designed to adhere to a conceptually simple ordering scheme which Randić assumes to be associated with branching, based on a binary label derived from a canonicalization of the adjacency matrix. In support of the suitability and utility of χ as a branching measure, Randić points out that it is usefully correlated with molecular properties such as the boiling point, and that it is minimal for a star graph and maximal for a path graph.[1] Despite the lack of a consensual definition of branching, these two aspects are also commonly found in the works of other authors. χ is also found to behave similarly to the Wiener index W [44] and Hosoya's Z [29], which Randić also considers theoretical indices of branching. Nonetheless χ has not been accepted as the one descriptor which defines branching, which is reflected in its renaming as connectivity index [43] in the scientific literature.

The Wiener index prominently appears in the papers on branching by Bonchev [6] and Bonchev et al. [8,9,7]. In [8], the authors numerically compare some existing indices (W, Z, χ and the largest eigenvalue of the adjacency matrix λ_1 [27]) and three newly introduced descriptors I_{pc}, I_D^E and I_W^E with relation to branching. A major part of the paper is concerned with the construction of a set of rules on branching (aided

[1] The *star graph* is the tree of order N with $N-1$ vertices of degree 1. The *path graph* is the tree of order N without vertices of a degree greater than 2.

by an analytical study of the behaviour of W and, to a lesser extent, I_D^E, both of which are assumed to decrease as the degree of branching increases), for example that

> [i]n a tree having a constant number of vertices, the branching increases when the number of branches attached to a given vertex increases owing to the chain decrease.

These rules are introduced in isolation from each other instead of within a unifying framework, which makes them potentially useful for a characterization of branching, but does not constitute a formal definition of the concept.

The theory was further developed and generalized in [6] to eliminate some of the strict preconditions for the applicability of the branching rules. It was taken up again in [7] with a focus on long-chain branching patters in polyolefins. In this work, the relative importance of the influencing factors on the overall degree of branching was found to be number of branches $\gg$ branch length $>$ branch centrality $\approx$ branch clustering $>$ total molecular weight $\approx$ backbone molecular weight, again based in a large part on the hypothesis that an increase of branching is followed by a decrease of the Wiener index.

Bertz [2] has been critical of attempts to create a model of branching from complex theories such as topological indices, calling these approaches forced fits. The rule-based characterization of branching by Bonchev et al., specifically [8], is subject to particular scrutiny. As an example for Bonchev and Trinajstić's reliance on a preconceived 'mathematical formalism' he highlights the targeted design of their branching rules to bring them into accordance with the information-theoretic measures I_D^E and I_D^W they have introduced. In Bertz' opinion, the criticized work illustrates the importance of letting the conceptual model dictate the mathematics and not vice versa.

Bertz instead offers two axioms and uses them to develop a mathematically grounded theory which does not rely on an absolute quantification of the degree of branching. He instead proposes a scheme which establishes a total order on the set of trees with regard to branching, which he considers superior to an earlier partial ordering scheme by Gutman and Randic [23] in which many structures remain 'un-comparable'. The strategy of this work is based on synthesis, the creation of complex systems from simple ones, which the author argues is a foundational concept in chemistry and should therefore be at the heart of defining 'the logical structure of chemistry'.

Perdih and Perdih [39] employ the statistical procedure of principal component analysis (PCA) to find a suitable reference property for the quantification of branching, in the form of either a molecular property or a topological descriptor. Their analysis was performed on five data sets, each containing between 18 and 40 trees representing alkanes. The axes of the reduced information spaces resulting from the PCA were used to find major influencing factors on branching in decreasing order of importance: the number of vertices, the number of branches, the existence of vertex degree of 3 versus 4, and finally the location of the branches.

Despite the vagueness of the concept of branching, it is possible to find consensus on some issues in the literature. For example, there is widespread agreement that some topological descriptors can indeed be used to quantify the degree of branching of a

tree to some extent, although not necessarily as the basis of an outright definition. For example, Kirby [30] notes that it is unclear whether a single branching index that is useful [can] be formulated [or whether] it should be accepted that several separately applied elements of branching may be necessary.

The requirement for branching indices to assign minimum and maximum values to the star and path graphs can be found frequently, for example in [8,19,40]. The idea that the number of branches is the most important facet of branching (at least if the number of vertices is constant) is also elementary enough to be found in several places, such as [7] and [39].

Whenever topological descriptors were analyzed numerically, the analysis was restricted to very small sets of graphs, containing e.g., 45 trees in [8] or 40 in [39]. Our own analysis uses much larger sets.

3 Methods

We have based our study on using exhaustive sets of trees of orders $N = 15$–20, which in total contain 1,340,577 non-isomorphic graphs. The majority of the analysis was performed on the 823,065 trees of order 20, while the other trees were mainly used to get an understanding of the influence of the number of vertices on the descriptor values. We have chosen to limit our investigation to trees because branching is a more elusive concept in the context of cyclic graphs. Exhaustive sets were used to avoid the issues of statistical significance and representatively generating random trees.

32 numeric values, 20 of them topological descriptors, were calculated for each of these graphs. We then processed the resulting data for each descriptor according to three criteria in order to identify which ones are associated with branching.

3.1 Generation of trees

The algorithm by Wright, Richmond, Odlyzko and McKay (WROM), originally given in [45], has been used to obtain the exhaustive sets of trees. It is based on an earlier algorithm for the generation of rooted trees by Beyer and Hedetniemi [3], whose underlying idea is to enumerate valid level sequences $(\ell_1, \ell_2, \ldots, \ell_N)$, where ℓ_i is the distance of the vertex i to the root, vertex 1 is the root ($\ell_1 = 0$) and two vertices i and j are adjacent if and only if $\ell_i + 1 = \ell_j$ and there is no k ($i < k < j$) such that $\ell_k \leq \ell_i$.

The WROM algorithm uses the same concept, but also defines a canonical level sequence representation for each free (unrooted) tree and skips all non-canonical items. It thus generates each non-isomorphic free tree of a specified order in an amortized runtime complexity of $O(1)$. Our implementation was written in the Python programming language as a generator function which yields NetworkX [26] objects.

3.2 Topological descriptors, elementary measures and Bertz ranks

The 20 descriptor values to be calculated for each tree were selected to capture different features of a graph (such as vertex degrees, distances and the eigenvalues of

the adjacency matrix) and to comprise both descriptors with some known association with branching as well as others. From the group of purported branching measures these were:

- χ, the Randić connectivity index [40],
- W, the Wiener index [44],
- I_D^E, the total information on distances [8],
- I_D^W, the total information on the realized distances [8],
- λ_1, the largest eigenvalue of the adjacency matrix [27],
- E, the graph energy [21], and
- B, the Bertz branching index [2].

B is characterized by very high degeneracy due to its very limited image of the integers in $[N-2, \frac{N^2-3N+2}{2}]$. For this reason, and motivated by its usage as the first element in the lexicographical ordering scheme suggested in [2], we have defined an extension that we will call the second-order Bertz branching index:

- $B_2(G) = B(G) + \frac{B(L(G))}{1+B(L(S_N))}$, where $L(G)$ denotes the line graph of G and S_N is the star graph of the same order as G.

The line graph for a given graph is created by representing each edge from the original graph as a vertex, and inserting an edge between two of these vertices if and only if the corresponding edges from the original graph share an incident vertex.

The remaining twelve descriptors do not have a reputation of sensitivity to branching. They are

- Z_2, the second Zagreb group index [25],
- MZI, the modified Zagreb index [37],
- VZI, the variable Zagreb index [37],
- AZI, the augmented Zagreb index [20],
- $\log PRS$, the logarithm of the product of the row sums of the distance matrix [41],
- $H_{A,2}$, the entropy with respect to the distribution of the eigenvalues of the adjacency matrix [17],
- I_{orb}^V, the topological information content [36],
- J, Balaban's J index [1],
- OdC, the offdiagonal complexity [11],
- I_{f^V}, the entropy based on the j-sphere functional [12],
- I_{f^C}, the entropy based on the vertex centrality functional [12], and
- I_{f^Δ}, the entropy based on the degree-degree association functional [16].

I_{f^V}, I_{f^C} and I_{f^Δ} are parametric measures. For the coefficients used in the functionals we have used an exponentially decreasing sequence $c_i = de^{i-1}$, where d is the diameter of the graph. Furthermore, our basis for I_{f^Δ} is $\alpha = 0.5$.

We also determined the ordering described by Bertz [2] for all trees, according to this rule: Pairs of [graphs] are ordered by comparing the sequences generated by counting the number of [edges] in the iterated line graphs so that the one which ultimately dominates is the more branched.

Then for each of our trees, BR is its position in the ordered list of all trees of the same order, such that BR is 1 for the least branched and maximal for the most branched tree according to Bertz.

Finally, the following elementary measures were used as readily interpretable quantities with a relationship to branching:

- N, the number of vertices,
- T, the number of terminal vertices, i.e., those with a degree of 1,
- $\delta_{\max}$, the maximum vertex degree,
- d, the diameter, i.e., the maximum distance between any two vertices,
- $dBB_{\min,\mathrm{avg},\max}$, the minimum, average and maximum distances between all pairs of branched vertices, and
- $dBC_{\min,\mathrm{avg},\max}$, the minimum, average and maximum distances between all branched vertices and the closest central vertex.

In these definitions, a *branched vertex* is a vertex of a degree greater than 2, and a *central vertex* is one whose eccentricity (the greatest distance to any other vertex in the graph) is minimal.

Again, a Python program was created and used to calculate these values. The code heavily uses the NetworkX library as well as the NumPy package [38] for efficient numerical computations. We originally planned to use the implementations of descriptors written in the R programming language from the QuACN package [35] for the task, but this implementation quickly turned out to be unacceptable in terms of computation time for the present application.

3.3 Analysis

As we have seen, there is no single, well-established objective criterion for what constitutes a good branching index. Our analysis is based on three main criteria, two of which employ Spearman's rank correlation coefficient ρ, whose definition was taken from [33]:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

where x_i and y_i are the ranks (the positions at which each raw value appears in the non-descending sequence of the raw values) of the studied variables, and $\bar{x}$ and $\bar{y}$ are the arithmetic means of these ranks.

The criteria are as follows:

1. Based on intuition and inspired by previous work by various authors (e.g., [8,40, 19]), we first demand that on a set of trees of the same order, a good branching index attains one extreme value (minimum or maximum) for the path graph and the opposite extremum for the star graph. We are less strict on the latter point because one could conceivably argue that a graph more structurally irregular than the star graph is in fact more highly branched, but failure to single out the path graph is surely not acceptable for a branching index.
2. We expect a branching index to be favorably correlated to the number of branches when the number of vertices is set to a constant. For this purpose the number

of branches is straightforwardly defined as the number of terminal vertices T, as proposed in [6]. We require at least $|\rho| \geq 0.7$, preferably $|\rho| \geq 0.8$, which can be interpreted as $\geq$49 or $\geq$64 %, respectively, of the variance of the ranked index values being explained by the ranked T values as per the coefficient of determination [42]. Visual examination of the relevant plots is also used to confirm that a useful relationship exists. This approach is also motivated by the literature, particularly [7] and [39], where the number of branches has been found to be the most important factor with an influence on branching.

3. On a more informal basis, the relationship between each descriptor and the Bertz ordering scheme will be investigated, with a strong correlation again pointing to a good measure of branching.

The necessary processing and visualization of the obtained data was done in the R statistical computing language.

4 Results

4.1 Computation

All computations were performed on a PC, based on an Intel i3 CPU with a clock speed of 2.1 GHz and 4 GiB of physical main memory, in a Debian GNU/Linux environment. The full run of generating the trees and calculating all the values listed in Sect. 3.2 has taken about 12 h of wall-clock time. This runtime could be drastically cut by utilizing multiple processor cores, as the code at hand does not perform any parallelization whatsoever.

The algorithm to produce the Bertz ordering scheme has turned out to be problematic. While the comparison between two trees usually comes to a decision within few iterations of line graphs, some pairs of trees require more steps and the computation of vast line graphs with millions of vertices. This led to paging and therefore significant computation times measured in tens of minutes.

4.2 Raw numerical data

A statistical overview of the descriptor values calculated on the sets of trees of the orders 15 and 20 can be found in Table 1.

4.3 Extreme value criterion

Our analysis of the raw data based on all the individual tree sets (orders 15–20) has shown that these indices do not take on a minimum or maximum for the path graph:

- I_{f^Δ}, although it has a maximum for the star graph.
- $H_{A,2}$, although it has a minimum for the star graph.
- OdC, although it has a minimum for the star graph.
- I^V_{orb}, although it has a minimum for the star graph. Its maximum value is shared by many different trees, but the path graph is not among them.

Table 1 Overview of the raw numerical data. N … number of vertices, $\bar{x}$ … arithmetic mean, s … standard deviation, ρ_T … rank correlation coefficient of descriptor and number of terminal vertices, ρ_{BR} … rank correlation coefficient of descriptor and rank in the Bertz ordering

	$N = 15$		$N = 20$			
Descriptor	$\bar{x}$	s	$\bar{x}$	s	ρ_T	ρ_{BR}
I_{fV}	3.81	0.02	4.22	0.02	−0.46	−0.52
I_{fC}	3.74	0.04	4.15	0.04	−0.91	−0.94
$I_{f\Delta}$	0.72	0.61	0.63	0.56	−0.33	−0.29
W	369.26	45.55	796.14	97.83	−0.74	−0.75
χ	6.67	0.36	8.91	0.40	−0.92	−0.94
B	21.75	5.25	30.18	5.95	0.89	1.00
B_2	21.81	5.29	30.21	5.97	0.88	1.00
λ_1	2.42	0.19	2.51	0.18	0.74	0.92
E	16.17	1.22	21.82	1.36	−0.85	−0.85
$H_{A,2}$	3.43	0.25	3.87	0.20	−0.80	−0.77
I_E^W	281.48	31.09	564.18	52.44	−0.71	−0.69
I_D^W	2410.79	284.63	5872.98	695.95	−0.75	−0.76
J	4.60	0.70	5.19	0.75	0.72	0.74
OdC	1.26	0.22	1.36	0.21	0.55	0.75
I_{orb}^V	3.33	0.39	3.81	0.31	−0.71	−0.76
$\log PRS$	83.69	2.68	125.46	3.54	−0.73	−0.74
Z_2	78.85	13.69	111.52	17.00	0.87	0.96
MZI	3.41	0.35	4.51	0.39	−0.86	−0.83
AZI	92.87	12.09	130.49	13.87	−0.44	−0.40
VZI	8.17	0.70	10.99	0.77	0.86	0.83

- The Zagreb group indices MZI, AZI and VZI, although they have a minimum for the star graph.

Because we consider it fundamental that any useful branching index singles out the path graph as the least branched tree by not assigning some intermediate value to it, we will not consider these to qualify.

Among the trees, the two functional-based entropy measures I_{fC} and I_{fV} are both maximal for the path graph, but they take on their minimum values for branched trees other than the star graph. We will keep this in mind since intuitively, the tree with the largest number of branches might be considered the most highly branched, but we do not immediately exclude them. The difficulty of analytically finding graphs for which these entropies become minimal has been acknowledged in [32]. Among all graphs, for example, I_{fV} (with an exponentially decreasing coefficient sequence) becomes maximal for a class of graphs called sphere-regular in the referenced paper.

The other indices studied, namely χ, W, I_D^E, I_D^W, λ_1, E, B, B_2, Z_2, $\log PRS$ and J, fully satisfy the criterion.

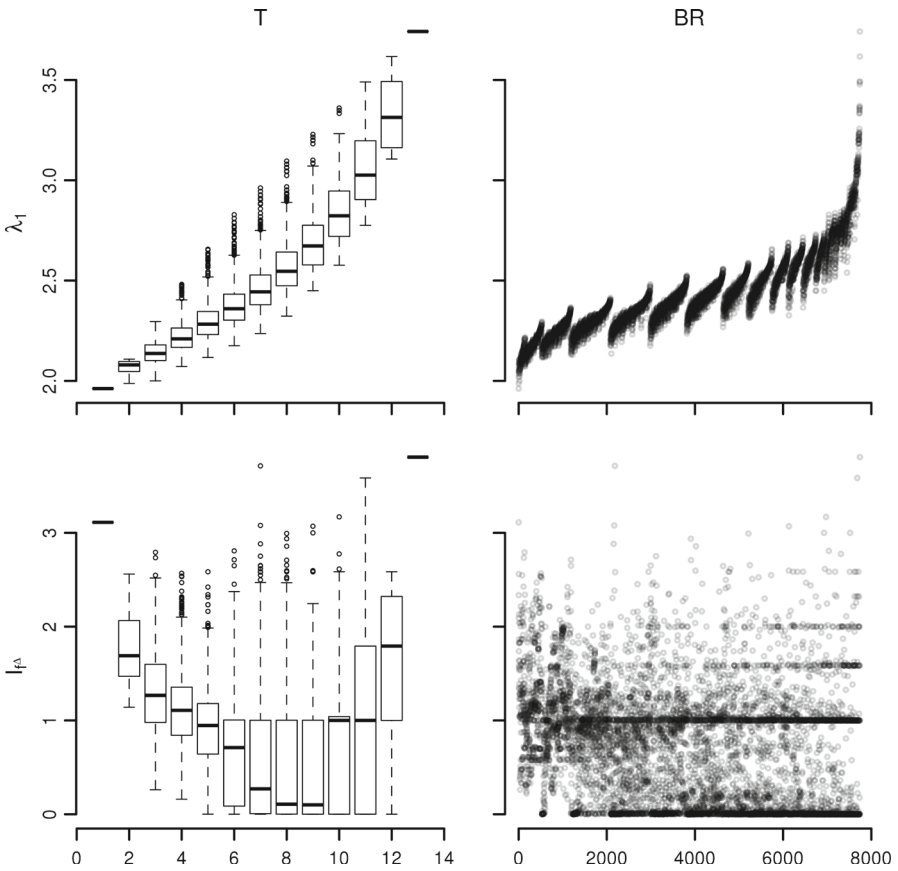

Fig. 1 *Left column Box plots* of descriptor values for trees with a given number of terminal vertices T. *Right column Scatter plot* of descriptor values against ranks in the Bertz ordering *BR*. The descriptor in the *top row* is the maximum eigenvalue λ_1, the one in the *bottom row* is the entropy based on the degree-degree association functional I_{f^Δ}. *Plots* based on the data for all trees of order 15

4.4 Number of branches criterion

The analysis in this section is based on rank correlation between each descriptor and the number of branches T (see Table 1), as well as plots such as those shown in the left column of Fig. 1.

Unsurprisingly, there is overlap between the indices failing the extreme value criterion and those with a particularly weak correlation to the number of branches: I_{f^Δ}, *OdC* and *AZI* all have $\rho < 0.6$ with regard to T in the case of $N = 20$. This suggests that these indices really do not usefully reflect the degree of branching of a tree, thereby reaffirming the decision to not consider them further. I_{f^V} can also be safely disqualified on that basis, at least with the chosen parametrization.

We also suggest that $H_{A,2}$ and I^V_{orb} can indeed be dropped based on the result from the previous section. Although a general trend to take on lower values as the number

of branches increases is recognizable for both of them, their relationships with the number of branches are characterized by many outliers (and an unusual shape of the correlation especially for $H_{A,2}$, whose distribution of values for groups of trees with the same number of branches behaves unpredictably in terms of average values and variance).

Only *MZI* and *VZI* would appear quite suitable according to this correlation criterion while not satisfying the extreme value criterion. We argue that it nonetheless remains sensible to leave them out, considering their shared property of becoming extreme for trees with an intermediate number of branches (e.g., a tree with $T = 8$ for the set of trees with $N = 20$), and the fact that Z_2 exhibits a similar distribution without this issue.

We will also disregard B and B_2. While they are correlated rather well to the number of terminal vertices, they are useless as standalone indices due to their degeneracy, with B taking on only 92 distinct values for the 823,065 trees of order 20. The B_2 index we have introduced experimentally fares only somewhat better with 4,083 unique values.

The ten remaining indices are all clearly correlated with the number of branches T. In the following, the elements of the partition of an exhaustive set of trees by T (i.e., the sets forming the columns in the boxplots above) will be referred to as *groups*. This is our interpretation of the findings:

- I_{fC} and χ both exhibit numerically very strong negative correlations with T, with $|\rho| \geq 0.9$ for $|V| = 20$. For I_{fC}, the per-group average values decrease almost linearly as the number of branches increases. There is relatively little overlap of values between groups for both I_{fC} and χ. It is interesting to note that their performance is similar despite their completely different definitions: χ is a degree-based, non-information-theoretic descriptor, whereas I_{fC} is an entropy measure built on a concept of vertex centrality [12].
- Z_2 is weaker in terms of its correlation coefficient with regard to T. As opposed to the preceding two indices, it tends to grow with the number of trees. Its variance is higher in groups of trees with many branches than it is for those with fewer; this is a characteristic it shares with χ. It is nonetheless capable of separating groups to an adequate degree.
- The spectral measure λ_1 is clearly positively correlated with the number of branches, but it does not separate the groups particularly well. E, which is also based on the eigenvalues of the adjacency matrix, exhibits a much stronger negative correlation.
- Neither of the three distance-based measures W, I_D^W and I_E^W, previously investigated by Bonchev et al. [8], has an outstandingly strong correlation to T. The same is true for the other two indices that are based on distances, namely J and $\log PRS$.

Supposedly, the main property through which the latter five descriptors are correlated to the number of branches are the shorter distances in a more highly branched tree. For example, the absolute values of the rank correlation coefficients ρ_d between each descriptor and the graph diameter are consistently much higher than those between each descriptor and the number of terminal vertices ρ_T for these indices ($|\rho_T|$ around 0.72 and $|\rho_d|$ around 0.92 for $N = 20$), while the opposite is true for the other indices which satisfy this criterion. Moreover, we have employed partial correlation to statistically remove the effect of d, using the partial rank correlation coefficient [33]

$$\rho' = \frac{\rho_{x,y} - \rho_{x,z}\rho_{y,z}}{\sqrt{\left(1 - {\rho_{x,z}}^2\right)\left(1 - {\rho_{y,z}}^2\right)}}$$

(measuring the correlation between two variables x and y when the effect of a third variable z has been removed). Indeed, this yields absolute values of ρ' no higher than 0.35 for the distance-based indices paired with T when the effect of d is removed. For comparison, $|\rho'|$ values for χ and I_{fC} are still above 0.8, even though these descriptors are clearly also linked with d.

We have also calculated partial correlation coefficients to study the association between the descriptors and other elementary measures (maximum degree $\delta_{\max}$, branch–branch distances $dBB_{\min,\mathrm{avg},\max}$ and branch–center distances $dBC_{\min,\mathrm{avg},\max}$). We hoped to find some obvious trends in this analysis, ideally such that there is little correlation between an index and an elementary measure on the whole, but a clear link when the effect of T is accounted for. Unfortunately, due to the strong variation of tree structures within groups of constant T, the data obtained fails to strongly support claims of this kind, although we could identify two trends: As the number of vertices and terminal vertices remains constant,

- decreases in branch–branch and branch–center distances, as well as
- increases in the maximum vertex degree

generally lead to a change of descriptor value in the same direction as when branches are added. In other words, if one were to attempt to derive a characterization of branching from a topological descriptor, it would be natural to argue that branching increases as branches move closer together.

4.5 Bertz ranking criterion

For the investigation with regard to this criterion, we have again used rank correlation, this time between each descriptor and the rank BR of the graph in the Bertz ordering of all trees of the same order; see the ρ_{BR} column in Table 1. The relevant scatter plots of these variables, two examples of which are shown in the right column of Fig. 1, were also considered.

For most of the descriptors, the ρ values when paired with the positions in the Bertz ranking BR are suggestive of a basic level of correlation with the positions in the Bertz ordering. Based on weak rank correlations and scrutiny of the plots, we claim that I_{fV}, $I_{f\Delta}$, $H_{A,2}$, OdC, I^V_{orb} and AZI are not linked to BR satisfactorily. $I_{f\Delta}$ is notable for its particularly chaotic-looking scatter plot (see the lower row of Fig. 1); this entropy measure is notoriously difficult to explain in terms of intuitive structural properties [31]. The perfect rank correlations for B and B_2 are of course unremarkable, as these descriptors form the basis for the initial two steps of the iterative ordering algorithm.

Z_2, I_{fC} and χ follow the ordering scheme especially closely. Conversely, the five distance-based descriptors W, I^W_E, I^W_D, J and $\log PRS$ deviate a lot from an ideal monotonic relationship, as do the remaining Zagreb group indices MZI and VZI. Out of the two spectral measures, λ_1 fares well, whereas E is less nicely associated.

4.6 Remarks on the performance on the full set

The correlation between descriptors and T/BR is far weaker on the entire set of trees of orders 15–20 than on the individual sets in which N is constant. B, B_2, λ_1, J and Z_2 can be said to somewhat retain their useful properties based on the ρ values and plots, while this is not at all the case for I_{fV} and χ. *VZI* also performs rather well in this context, which is an interesting case in that we have not considered it as a potential branching index due to its failure to single out the path graph and despite its comparatively high correlation coefficient. The rest of the descriptors, including the distance-based indices other than J, cannot be said to be associated with T in the set of all trees.

The bad performance of the descriptors on the full set is caused by the different ranges of the values of some descriptors for trees of different orders, which can be seen to some extent in Table 1. Due to this property of many indices, we think that normalization might be helpful to achieve better correlations in some cases. Especially for the descriptors which have their extrema for the path and star graph, it should not be difficult to re-scale their values to a common range such as [0, 1]. Whether or not this leads to more desirable results is out of scope for this study.

5 Summary and conclusion

Based on our analysis, we have newly discovered a branching index, namely the entropy based on the vertex centrality functional I_{fC}, see [12]. To our knowledge, the capability of this descriptor to capture branching has not been assessed previously. Although I_{fC} corresponds favorably to the number of terminal vertices, the distances between branches and the Bertz ordering, its failure in singling out the star graph as the most branched tree has to be noted as a major shortcoming. It would be interesting to see how the behavior of the index would change if its parameters were chosen differently.

We have also found that χ fits our criteria very well, which ties in with Randić's original introduction of the descriptor as a branching index [40], even though it is usually referred to as connectivity index in the scientific literature. The second Zagreb group index, denoted here by Z_2, has a definition very much like χ. This causes these two descriptors to behave similarly, as already noted in [24]. Interestingly, Gutman et al. do not consider Z_2 to be a branching measure in this paper based on their numerical analysis. This might be due to their restriction to trees with a maximum degree of 3, but it is certainly also demonstrative of the arbitrariness of the many attempts to quantify branching.

The basic suitability of W, I_E^W and I_D^W as branching indices can also be confirmed by our analysis, although they do not stand out in a way that would warrant using them as a basis for a definition of branching (as Bonchev et al. have done). Indeed, their performance is very similar to that of two other distance-based measures, J and $\log PRS$. We have seen that the main mechanism through which these descriptors capture branching according to our criteria are the reduced overall distances in the tree, and that they do not respond consistently to the number of terminal vertices beyond this aspect.

The widespread acceptance of the maximum eigenvalue λ_1 as a branching index (see, for example, [8,30]) also is not contradicted by our analysis. On the other hand, the second spectral descriptor we have studied in this work, E, does not appear to be typically seen this way. It is probably less useful than λ_1 (an assumption we base both on its irregular behavior visible in the plots against the number of terminal vertices and the ranks in the Bertz ordering scheme), but we do think that it can be called a branching index nonetheless.

The descriptors that we have not mentioned in this section do not seem to be suitable as branching measures. Furthermore, most of the indices we did identify as possible branching measures break down when one tries to use them to compare trees of different orders. λ_1, J and Z_2 are exceptions to this rule.

The investigation of the relationship between topological indices and the Bertz ordering scheme has been a new contribution of this work. We have found the ordering scheme to be an interesting idea with limited practical applicability due to its computational cost. This makes the correlation between the ranks in the ordering and some of the numerical descriptors potentially interesting.

Based on the differences between descriptors that are seen as branching measures, the long-standing elusiveness of a precise definition of branching and the subjectivity of the concept, we agree with Kirby [30] that it is dubious whether branching *should* be considered as a single unified concept. It is probably more sensible to primarily consider the individual features which can easily be quantified, such as the number of and distances between branches, in terms of a multivariate concept of branching, in a way that makes sense for the respective task. We nevertheless do not want to deny the utility of topological descriptors in this context, as some of them clearly are linked to branching in a major way, and it is usually more convenient to deal with a single scalar value than with an entire system of interrelated properties.

6 Discussion

We have used three different methods of assessment to examine topological descriptors with respect to branching, namely the extreme value criterion, the correlations with elementary measures (based on the entire exhaustive sets as well as some pattern-based sets of trees) and the association with the Bertz ordering. These approaches do not have an immediately obvious common basis. Thus, it is reassuring that they are compatible insofar as they have yielded similar results. We consider this a strength in light of the lack of an accepted definition of branching, because the results can be seen as mutually supportive of each other.

Furthermore, we believe that all three criteria individually constitute sensible characterizations of branching. The extreme value criterion has been formulated by many different authors and reflects an immediately plausible notion. The correlation with elementary measures is also based on a widespread idea but allows for more fine-grained statements on the branching behavior of a descriptor. And while it is not as established, the Bertz scheme has the useful property of being a total ordering, an advantageous trait when used in a correlation analysis, while fitting in nicely with the other two criteria.

The range of the studied trees is an interesting issue. On the one hand, trees containing vertex degrees higher than 4 do not correspond to chemical structures, but we have nonetheless included them in our analysis. A comparison to results for a narrower set of only chemical trees might give rise to insightful conclusions. On the other hand, it would also be worthwhile to include cyclic structures, as far less work has been done so far on the branching of cyclic graphs.

Regardless of the underlying set of graphs, one must take care not to overgeneralize the results. It is natural to expect that the general behavior of the indices will likely be comparable for trees of lower and higher orders, but since we have not given any formal proofs, this cannot be guaranteed. It might seem unlikely for other tree sets to yield significantly different results, but such a finding would not be unprecedented. For an illustrative example of a surprising dependence on the number of vertices, see [22], in which Gutman et al. show that the atom-bond connectivity index has a minimum for the path graph for $4 \leq |V| < 10$, but not for $|V| \geq 10$. Instead of repeating the analyses from this work with different sets of trees, it would therefore be more useful to study why and how the individual indices reflect branching in the way they do, so that universally valid statements can be formulated.

Acknowledgments Matthias Dehmer thanks the Austrian Science Funds (project P22029-N13) and the Standortagentur Tirol for supporting this work.

References

1. A. Balaban, Highly discriminating distance-based topological index. Chem. Phys. Lett. **89**, 399–404 (1982)
2. S. Bertz, Branching in graphs and molecules. Discret. Appl. Math. **19**, 65–83 (1988)
3. T. Beyer, S. Hedetniemi, Constant time generation of rooted trees. SIAM J. Comput. **9**, 706–712 (1980)
4. D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures* (Research Studies Press, Chichester, 1983)
5. D. Bonchev, D.H. Rouvray, *Complexity in Chemistry, Biology, and Ecology. Mathematical and Computational Chemistry* (Springer, New York, 2005)
6. D. Bonchev, Topological order in molecules 1. Molecular branching revisited. J. Mol. Struct. (Theochem) **336**, 137–156 (1995)
7. D. Bonchev, E. Markel, A. Dekmezian, Topological analysis of long-chain branching patterns in polyolefins. J. Chem. Inf. Comput. Sci. **51**(5), 1274–1285 (2001)
8. D. Bonchev, N. Trinajstic, Information theory, distance matrix, and molecular branching. J. Chem. Phys. **67**(10), 4517–4533 (1977)
9. D. Bonchev, N. Trinajstic, On topological characterization of molecular branching. Int. J. Quant. Chem. **12**, 293–303 (1978)
10. G. Chartrand, *Introductory Graph Theory* (Dover, New York, NY, 1985)
11. J. Claussen, Offdiagonal complexity: a computationally quick complexity measure for graphs and networks. Phys. Stat. Mech. Appl. **375**(1), 365–373 (2007)
12. M. Dehmer, Information processing in complex networks: graph entropy and information functionals. J. Appl. Math. Comput. **201**, 82–94 (2008)
13. M. Dehmer (ed.), *Structural Analysis of Complex Networks* (Birkhäuser, Cambridge, 2010)
14. M. Dehmer, Information theory of networks. Symmetry **3**, 767–779 (2012)
15. M. Dehmer, F. Emmert-Streib, T. Gesell, A comparative analysis of multidimensional features of objects resembling sets of graphs. Appl. Math. Comput. **196**, 221–235 (2008)
16. M. Dehmer, F. Emmert-Streib, Y. Tsoy, K. Varmuza, Quantifying structural complexity of graphs: information measures in mathematical chemistry, in *Quantum Frontiers of Atoms and Molecules*, ed. by M. Putz (Nova Publishing, New York, NY, 2011), pp. 479–498

17. M. Dehmer, L. Sivakumar, K. Varmuza, Uniquely discriminating molecular structures using eigenvalue-based descriptors. Match Commun. Math. Comput. Chem. **67**(1), 147–172 (2012)
18. F. Emmert-Streib, M. Dehmer, Networks for systems biology: conceptual connection of data and function. IET Syst. Biol. **5**, 185–207 (2011)
19. M. Fischermann, I. Gutman, A. Hoffmann, D. Rautenbach, D. Vidovic, L. Volkmann, Extremal chemical trees. Z. Naturforsch. **57a**, 49–52 (2002)
20. B. Furtula, A. Graovac, D. Vukicevic, Augmented Zagreb index. J. Math. Chem. **48**(2), 370–380 (2010)
21. I. Gutman, The energy of a graph. Ber. Math. Statist. Sekt. Forsch. Graz **103**, 1–22 (1978)
22. I. Gutman, B. Furtula, M. Ivanovic, Notes on trees with minimal atom-bond connectivity index. Match Commun. Math. Comput. Chem. **67**, 467–482 (2012)
23. I. Gutman, M. Randic, Algebraic characterization of skeletal branching. Chem. Phys. Lett. **47**(1), 15–19 (1977)
24. I. Gutman, B. Ruščić, T. Trinajstic, C. Wilcox Jr, Graph theory and molecular orbitals. XII. Acyclic polyenes. J. Chem. Phys. **62**(9), 3399–3405 (1975)
25. I. Gutman, N. Trinajstic, Graph theory and molecular orbitals. total φ-electron energy of alternant hydrocarbons. Chem. Phys. Lett. **17**(4), 535–538 (1972)
26. A. Hagberg, D. Schult, P. Swart, Exploring network structure, dynamics and function using NetworkX. In: G. Varoquaux, T. Vaught, J. Millman (eds.) Proceedings of the 7th python in science conference (SciPy2008) (Pasadena, CA, 2008), pp. 11–15
27. G. Hall, Eigenvalues of molecular graphs. Bull. Inst. Math. Appl. **17**, 70–72 (1981)
28. F. Harary, *Graph Theory* (Addison-Wesley Publishing Company, Reading, MA, 1969)
29. H. Hosoya, Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull. Chem. Soc. Jpn. **44**, 2332–2339 (1971)
30. E. Kirby, Sensitivity of topological indices to methyl group branching in octanes and azulenes, or what does a topological index index? J. Chem. Inf. Comput. Sci. **34**, 1030–1035 (1994)
31. V. Kraus, M. Dehmer, F. Emmert-Streib, Probabilistic inequalities for evaluating structural network measures (2013). Submitted for publication
32. V. Kraus, M. Dehmer, M. Schutte, On sphere-regular graphs and the extremality of information-theoretic network measures (2013). Accepted for publication
33. J. Lozán, H. Kausch, *Angewandte Statistik für Naturwissenschaftler*, 4th edn. (Wissenschaftliche Auswertungen, Hamburg, 2007)
34. A. Mehler, Social ontologies as generalized nearly acyclic directed graphs: a quantitative graph model of social tagging, in *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, ed. by M. Dehmer, F. Emmert-Streib, A. Mehler (Birkhäuser, Boston/Basel, 2011), pp. 259–319
35. L. Müller, K. Kugler, A. Dander, A. Graber, M. Dehmer, QuACN: an R package for analyzing complex biological networks quantitatively. Bioinformatics 27(1), 140–141 (2011). http://cran.r-project.org/web/packages/QuACN/
36. A. Mowshowitz, Entropy and the complexity of graphs: I. An index of the relative complexity of a graph. Bull. Math. Biophys. **30**(1), 175–204 (1968)
37. S. Nikolic, G. Kovacevic, A. Milicevic, N. Trinajstic, The Zagreb indices 30 years after. Croat. Chem. Acta **76**(2), 113–124 (2003)
38. T. Oliphant, Python for scientific computing. Comput. Sci. Eng. **90**, 9 (2007)
39. A. Perdih, M. Perdih, On topological indices indicating branching Part I. The principal component analysis of alkane properties and indices. Acta Chim. Slov. **47**, 231–259 (2000)
40. M. Randic, On characterization of molecular branching. J. Am. Chem. Soc. **97**(23), 6609–6615 (1975)
41. H. Schultz, E. Schultz, T. Schultz, Topological organic chemistry. 4. Graph theory, matrix permanents, and topological indices of alkanes. J. Chem. Inf. Comput. Sci. **32**(1), 69–72 (1992)
42. R. Taylor, Interpretation of the correlation coefficient: a basic review. J. Diagn. Med. Sonogr. **1**, 35–39 (1990)
43. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics, Second, Revised and Enlarged edn. Methods and Principles in Medicinal Chemistry* (Wiley, Weinheim, 2009)
44. H. Wiener, Structural determination of paraffin boiling points. J. Am. Chem. Soc. **1**(69), 17–20 (1947)
45. R. Wright, B. Richmond, A. Odlyzko, B. McKay, Constant time generation of free trees. SIAM J. Comput. **15**, 540–548 (1986)