

ON THE INFORMATION CONTENT OF WEIGHTED CHEMICAL STRUCTURES

MATTHIAS DEHMER AND STEPHAN BORGERT

ABSTRACT. To characterize molecule structures by using information-theoretic techniques is an interesting and challenging problem in mathematical chemistry. However, most of the classical information indices are only defined for characterizing the skeletons of chemical graphs that correspond to unweighted graphs. This work presents a possible extension of an information measure that was recently developed to determine the structural information content of unweighted graphs. The novel measure then takes edge and vertex labels into account when measuring the information content of a weighted network.

KEYWORDS: *Information Measures, Complex Systems, Structural Complexity, Labeled Chemical Graphs, Information Theory, Chemical Graph Theory*

2000 Mathematics Subject Classification: 68R10, 94C30

1 Introduction

A major part of the research in chemical graph theory deals with investigating structural properties of molecules representing graphs by using so-called topological and information-theoretic indices [3, 2, 12, 13]. The latter is the object of research we cover in the present paper. Generally, information theory [14] offers quantitative methods to investigate, e.g., information processing and information transmission in graphs [2]. Particularly in mathematical chemistry, many problems in QSAR and QSPR require methods for analyzing structural properties of chemical graphs quantitatively. QSAR (Quantitative structure-activity relationship) deals with describing pharmacokinetic processes as well

as biological activity or chemical reactivity [1, 8]. In contrast, QSPR (Quantitative Structure-Property Relationship) addresses the problem to convert chemical structures into molecular descriptors [16] which are relevant to a physico-chemical property or a biological activity [8, 9]. Especially, a main problem in QSPR is to investigate relationships between molecular structure and physicochemical properties [2].

In [6], we recently introduced a novel information measure to determine the structural information content of unlabeled and undirected chemical graphs (see Section (2)). So far, this entropic measure has been mainly used to detect molecular branching in molecules representing the just mentioned graph class [7].

As a main contribution of our paper, we want to extend the information measure presented in [7] to vertex- and edge-labeled (weighted) graphs because chemical structures can be adequately represented by graphs only if different types of atoms (vertices) and different types of bonds (edges) are considered. For instance, heteroatoms like nitrogen (N) or oxygen (O) instead of carbon (C) change many properties of the molecules considerably. Double bonds and triple bonds are much more reactive than single bonds.

A classical contribution to quantify the amount of information for the kind of atoms in a molecule was given by [2, 5]. By using the total or mean information [2], it turned out that the so-called information indices on atomic composition for molecules denoted by the empirical formulas $A_xB_yC_z$ can be determined as [2, 5]

$$I_{AC}^t := (x + y + z) \log(x + y + z) - x \log(x) - y \log(y) - z \log(z), \quad (1)$$

or

$$I_{AC}^m := -p_x \log(p_x) - p_y \log(p_y) - p_z \log(p_z), \quad (2)$$

where

$$p_x := \frac{x}{x + y + z}, \quad p_y := \frac{y}{x + y + z}, \quad \text{and} \quad p_z := \frac{z}{x + y + z}. \quad (3)$$

However, we see that these indices do not take into account any structural properties of the molecules. In contrast, in the following we put the emphasis on developing an information measure for characterizing chemical structures representing vertex- and edge-labeled graphs.

2 Structural Information Content of Unweighted Graphs

First, we briefly introduce mathematical preliminaries [4, 11, 10] to formulate our approach. For this, we first start with the definition of finite, undirected and connected graphs. We call $G = (V, E), |V| < \infty, E \subseteq \binom{V}{2}$ a finite, undirected and connected graph whereas \mathcal{G}_{UC} denotes the set of such graphs. In the following, we also repeat the definitions of some metrical properties of graphs [7, 15]. $d(u, v)$ denotes the shortest distance between $u \in V$ and $v \in V$. $d(u, v)$ is an integer metric. For $G \in \mathcal{G}_{UC}$, $\sigma(v) = \max_{u \in V} d(u, v)$ is called the eccentricity of $v \in V$ and $\rho(G) = \max_{v \in V} \sigma(v)$ the diameter of G , respectively.

$$S_j(v_i, G) := \{v \in V \mid d(v_i, v) = j, j \geq 1\}, \quad (4)$$

denotes the j -sphere of v_i regarding $G \in \mathcal{G}_{UC}$. Further, we state the following definitions.

Definition 2.1 *Let*

$$A_V^G := \{l_v^1, l_v^2, \dots, l_v^{|A_V^G|}\}, \quad (5)$$

and

$$A_E^G := \{l_e^1, l_e^2, \dots, l_e^{|A_E^G|}\}, \quad (6)$$

be unique (finite) vertex and edge alphabets, respectively. $l_V : V \rightarrow A_V^G$ and $l_E : E \rightarrow A_E^G$ are the corresponding edge and vertex labeling functions. Then, we call $G := (V, E, l_V, l_E)$ a finite, undirected and labeled graph. \mathcal{G}_{UL} denotes the set of finite, undirected and labeled graphs.

Definition 2.2 *Let X be a discrete random variable with alphabet A and $p(x_i) = Pr(X = x_i)$ be the probability mass function of X . Then, the entropy of X is defined by*

$$H(X) := - \sum_{x_i \in A} p(x_i) \log(p(x_i)). \quad (7)$$

Now we briefly repeat the construction of the entropy measure for determining the structural information content of unlabeled graphs that was recently introduced in [7, 6]. In [7], the resulting information measure was mainly used to detect molecular branching in unlabeled chemical graphs. As mentioned, the chemical graphs were considered as skeletons only, i.e., all atoms

and all bonds were considered as equal. To define a probability distribution by inferring structural characteristics from the graphs under consideration, we used the definition of the j -spheres, see Equation (4). Finally, we defined the quantities

$$p^V(v_i) := \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}, \quad (8)$$

which have been interpreted as vertex probabilities [7]. Further,

$$f^V(v_i) := \alpha^{c_1|S_1(v_i,G)|+c_2|S_2(v_i,G)|+\dots+c_\rho(G)|S_\rho(G)(v_i,G)|}, \quad c_k > 0, 1 \leq k \leq \rho(G), \alpha > 0, \quad (9)$$

represents an information functional that captures structural information of a graph. c_k are arbitrary real positive coefficients to be chosen such that they are not all equal. We see that this information functional is based on the inferred j -sphere cardinalities. From these definitions, the structural information content of $G \in \mathcal{G}_{UC}$ has been defined as its corresponding entropy [6]. As a result, we obtained

$$I_{f^V}(G) := - \sum_{i=1}^{|V|} \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)} \log \left(\frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)} \right), \quad (10)$$

that represents a family of graph entropy measures. From a mathematical viewpoint, it turned out [7] that this entropy measure generalizes most of the classical information indices used in mathematical and computational chemistry [2, 16].

3 Structural Information Content of Weighted Graphs

In this section, we want to extend the proposed entropy measure from the previous section for determining the structural information content of vertex- and edge-labeled graphs. Let $G = (V, E) \in \mathcal{G}_{UL}$ and let $v_i \in V$ be an arbitrary vertex. Further, we assume that G is arbitrarily labeled (regarding the labels

$v_i, 1 \leq i \leq |V|$). In the following, we define the paths

$$P_1^j := (v_i, v_{a_1}, v_{a_2}, \dots, v_{a_j}), \quad (11)$$

$$P_2^j := (v_i, v_{b_1}, v_{b_2}, \dots, v_{b_j}), \quad (12)$$

\vdots

$$P_{k_j}^j := (v_i, v_{x_1}, v_{x_2}, \dots, v_{x_j}), \quad (13)$$

which are induced by determining the shortest paths starting from v_i . Hence, the corresponding edge sets are given by

$$E_1^j := \{\{v_i, v_{a_1}\}, \{v_{a_1}, v_{a_2}\}, \dots, \{v_{a_{j-1}}, v_{a_j}\}\}, \quad (14)$$

$$E_2^j := \{\{v_i, v_{b_1}\}, \{v_{b_1}, v_{b_2}\}, \dots, \{v_{b_{j-1}}, v_{b_j}\}\}, \quad (15)$$

\vdots

$$E_{k_j}^j := \{\{v_i, v_{x_1}\}, \{v_{x_1}, v_{x_2}\}, \dots, \{v_{x_{j-1}}, v_{x_j}\}\}, \quad (16)$$

and we set

$$E^j := E_1^j \cup E_2^j \cup \dots \cup E_{k_j}^j. \quad (17)$$

Now, for $j = 1, 2, \dots, \rho(G)$, we define the quantities

$$|S_j^\mu(v_i, G)| := |\{v \in V \mid d(v_i, v) = j, j \geq 1, l_V(v) = l_v^\mu, \mu = 1, 2, \dots, |A_V^G|\}|, \quad (18)$$

and

$$|E_{l_e^\mu}^j| := |\{e \in E_1^j \cup E_2^j \cup \dots \cup E_{k_j}^j \mid e \text{ is incident with } v \in S_j(v_i, G) \wedge l_E(e) = l_e^\mu, \mu = 1, 2, \dots, |A_E^G|\}|. \quad (19)$$

To illustrate the given definitions, we consider Figure (1). This figure shows a special chemical structure represented by a vertex- and edge-labeled chemical graph G . We set $A_V^G = \{O, C, N\}$ and $A_E^G = \{s, d\}$. O, C and N denote the kind of atoms. The edge type s represents a single bond whereas d represents a double bond. For example, if we choose v_3 as a starting vertex, we yield

$$|S_1^O(v_3, G)| = 1, |S_1^C(v_3, G)| = 1, |S_1^N(v_3, G)| = 1, \quad (20)$$

$$|S_2^O(v_3, G)| = 0, |S_2^C(v_3, G)| = 2, |S_2^N(v_3, G)| = 0, \quad (21)$$

$$|S_2^O(v_3, G)| = 1, |S_2^C(v_3, G)| = 0, |S_2^N(v_3, G)| = 0, \quad (22)$$

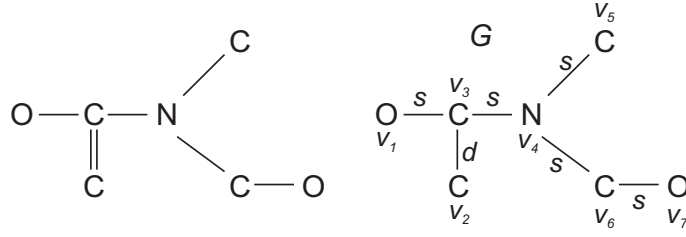


Figure 1: Left: A chemical structure representing C₄H₉NO₂. Right: The corresponding vertex- and edge-labeled graph G .

and

$$|E_s^1| = 2, |E_d^1| = 1, \quad (23)$$

$$|E_s^2| = 2, |E_d^2| = 0, \quad (24)$$

$$|E_s^3| = 1, |E_d^3| = 0. \quad (25)$$

By using the above given definitions, we first obtain direct generalizations of the information functional expressed by Equation (9).

Definition 3.1 Let $G = (V, E) \in \mathcal{G}_{UL}$ and we assume that $A_V^G \neq \emptyset$. We define

$$f_l^V(v_i) := \alpha^{\sum_{k=1}^{\rho(G)} \sum_{\mu=1}^{|A_V^G|} c_k^{\mu} |S_k^{\mu}(v_i, G)|}, c_k^{\mu} > 0, \alpha > 0. \quad (26)$$

Definition 3.2 Let $G = (V, E) \in \mathcal{G}_{UL}$ and we assume that $A_E^G \neq \emptyset$. We define

$$f_l^E(v_i) := \alpha^{\sum_{k=1}^{\rho(G)} \sum_{\mu=1}^{|A_E^G|} b_k^{\mu} |E_{i_e}^{\mu}|}, b_k^{\mu} > 0, \alpha > 0. \quad (27)$$

Now, it is straightforward to derive an information functional for quantifying structural information of vertex- and edge-labeled graphs.

Definition 3.3 Let $G = (V, E) \in \mathcal{G}_{UL}$ and we assume that $A_V^G, A_E^G \neq \emptyset$. We define

$$f_l^{V,E}(v_i) := \alpha^{\sum_{k=1}^{\rho(G)} \sum_{\mu=1}^{|A_V^G|} c_k^{\mu} |S_k^{\mu}(v_i, G)| + \sum_{k=1}^{\rho(G)} \sum_{\mu=1}^{|A_E^G|} b_k^{\mu} |E_{i_e}^{\mu}|}, b_k^{\mu}, c_k^{\mu} > 0, \alpha > 0. \quad (28)$$

From this, we yield the corresponding families of entropic measures to determine the structural information content of labeled graphs.

Definition 3.4 Let $G = (V, E) \in \mathcal{G}_{UL}$ and we assume that $A_V^G, A_E^G \neq \emptyset$. We obtain the following information measures:

$$I_{f_l^V}(G) := - \sum_{i=1}^{|V|} \frac{f_l^V(v_i)}{\sum_{j=1}^{|V|} f_l^V(v_j)} \log \left(\frac{f_l^V(v_i)}{\sum_{j=1}^{|V|} f_l^V(v_j)} \right), \quad (29)$$

$$I_{f_l^E}(G) := - \sum_{i=1}^{|V|} \frac{f_l^E(v_i)}{\sum_{j=1}^{|V|} f_l^E(v_j)} \log \left(\frac{f_l^E(v_i)}{\sum_{j=1}^{|V|} f_l^E(v_j)} \right), \quad (30)$$

$$I_{f_l^{V,E}}(G) := - \sum_{i=1}^{|V|} \frac{f_l^{V,E}(v_i)}{\sum_{j=1}^{|V|} f_l^{V,E}(v_j)} \log \left(\frac{f_l^{V,E}(v_i)}{\sum_{j=1}^{|V|} f_l^{V,E}(v_j)} \right). \quad (31)$$

References

- [1] R. Benigni. *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*. CRC Press, 2003.
- [2] D. Bonchev. *Information Theoretic Indices for Characterization of Chemical Structures*. Research Studies Press, Chichester, 1983.
- [3] D. Bonchev and N. Trinajstić. Information theory, distance matrix and molecular branching. *Journal of Chemical Physics*, 67:4517–4533, 1977.
- [4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley & Sons, 2006.
- [5] S. M. Dancoff and H. Quastler. Information content and error rate of living things. In H. Quastler, editor, *Essays on the Use of Information Theory in Biology*, pages 263–274. University of Illinois Press, 1953.
- [6] M. Dehmer. A novel method for measuring the structural information content of networks. *Cybernetics and Systems*, 39:825–843, 2008.
- [7] M. Dehmer and F. Emmert-Streib. Structural information content of chemical networks. *Zeitschrift für Naturforschung, Part A*, 63a:155–159, 2008.

- [8] J. Devillers and A. T. Balaban. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers, 1999. Amsterdam, The Netherlands.
- [9] M. V. Diudea. *QSPR / QSAR Studies by Molecular Descriptors*. Nova Publishing, 2001.
- [10] R. Halin. *Graphentheorie*. Akademie Verlag, 1989. Berlin, Germany.
- [11] F. Harary. *Graph Theory*. Addison Wesley Publishing Company, 1969. Reading, MA, USA.
- [12] E. V. Konstantinova. On some applications of information indices in chemical graph theory. In R. Ahlswede, L. Bäumer, N. Cai, H. Aydinian, V. Blinovskiy, C. Deppe, and H. Mashurian, editors, *General Theory of Information Transfer and Combinatorics*, Lecture Notes of Computer Science, pages 831–852. Springer, 2006.
- [13] M. Randić and D. Plavšić. On the concept of molecular complexity. *Croatica Chemica Acta*, 75:107–116, 2002.
- [14] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1997. Urbana, IL, USA.
- [15] V. A. Skorobogatov and A. A. Dobrynin. Metrical analysis of graphs. *Commun. Math. Comp. Chem.*, 23:105–155, 1988.
- [16] R. Todeschini, V. Consonni, and R. Mannhold. *Handbook of Molecular Descriptors*. Wiley-VCH, 2002. Weinheim, Germany.

Matthias Dehmer, Discrete Mathematics and Geometry, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria,
email: mdehmer@dmg.tuwien.ac.at

Stephan Borgert, Darmstadt University of Technology, Hochschulstr. 10, 64289 Darmstadt, Germany,
email: borgert@tk.informatik.tu-darmstadt.de