# Chapter 1

# Complexity and Diversity of Digraphs

**Steven H. Bertz**
Complexity Study Center
Mendham, NJ 07945 USA
sbertz@complexitystudycenter.org

**Gil Z. Pereira**
Hunter College High School
New York, NY 10128 USA

**Christina M.D. Zamfirescu**
Hunter College and Graduate Center–CUNY
695 Park Avenue, New York, NY 10021 USA

## 1. Introduction

The motto of the Complexity Study Center is "If it isn't complex, it isn't interesting." There has been a great deal of ferment in 'Complexity Science' in recent years, as chronicled in the proceedings of the International Conference on Complex Systems [Bar-Yam & Minai 2003, Bar-Yam 2000] and those of the Santa Fe Institute [Nadel & Stein 1995, Cowan 1994]. We have been primarily focused on developing metrics of complexity relevant to chemistry, especially synthetic chemistry [Bertz 2003a-c]. Our approach is to abstract a system, e.g., a molecule or a plan for its synthesis, as a graph and then to use the tools of graph theory to characterize the complexity and diversity of the system.

For the finer points of graph theory, we recommend standard texts [Gibbons 1991, Harary 1969] and limit our introduction to the basic definitions needed to appreciate the results. A *graph G* consists of a finite set $V(G)$ of *vertices* (or *points*) together with a set $E(G)$ of *edges* (or *lines*), which are unordered pairs of distinct vertices of $V(G)$. Mathematicians generally use the vertex-edge convention; however, following Harary [1969], authors in other fields usually use the point-line system, which we adopt here. A line $x = p_1p_2 = p_2p_1$ in *G joins* points $p_1$ and $p_2$, which are *adjacent points*. Two lines that share a point are *adjacent lines*, e.g., $p_1p_2$ and $p_2p_3$. Point $p$ and line $x$ are *incident* to each other. In a *multigraph* more than one line joins at least one pair of points.

Complexity and Diversity of Digraphs

When the lines of a graph are directed, i.e., ordered pairs of distinct points, they are called *arcs*, and the graph is called a *digraph* (*directed graph*) $D$. Thus, $p_ip_j$ is the arc from point $p_i$, the *tail*, to point $p_j$, the *head*. *Multidigraphs* allow multiple arcs, e.g., two arcs from $p_1$ to $p_2$. The *in-degree of* $p_i$, $in(p_i)$, is the number of arcs terminating at $p_i$, and the *out-degree of* $p_i$, $out(p_i)$, is the number of arcs originating from it. Since each arc has one head and one tail, for any digraph $\sum_i in(p_i) = \sum_i out(p_i)$. The *degree* $d_i$ of point $p_i$ in a graph or digraph is the number of lines incident to it, and in a digraph $d_i = in(p_i) + out(p_i)$.

A *walk* of length $n - 1$ is a sequence of points $p_1, p_2, p_3, \ldots, p_n$ that are joined by arcs $p_1p_2$, $p_2p_3, \ldots, p_{n-1}p_n$. In a *path* $P_n$ on $n$ points, each point and hence each line is distinct. In a *pseudopath* one or more of the arcs are oriented in the opposite direction from the rest. *Semipaths* are comprised of paths and pseudopaths. In a *connected* (di)graph all pairs of points are the endpoints of some (semi)path. A *cycle* $C_n$, also called an $n$-cycle or $n$-ring, is a sequence of arcs $p_1p_2, p_2p_3, \ldots, p_{n-1}p_n, p_np_1$ such that all $n$ points are distinct. We include $p_1p_2p_1$ as a 2-cycle. A *tree* is a connected graph without cycles.

Two graphs $G$ and $H$ are *isomorphic*, $G \cong H$, if and only if there exists a one-to-one correspondence between their point sets that preserves adjacency. An *invariant* of graph $G$ is a number $I(G)$ associated with $G$ that has the same value for any graph $H$ isomorphic to $G$. For example, the number of points, the number of lines and the number of pairs of adjacent lines are graph invariants. A *subgraph* $S$ of graph $G$ is a graph that has all its points in $V(G)$ and lines in $E(G)$. We include $G$ itself and also $P_1$, the trivial path on 1 point, in the set of all possible subgraphs of $G$. A *spanning subgraph* is a subgraph containing all the points of $G$. A spanning subgraph that is also a tree is a *spanning tree*.

A molecule can be abstracted as a *molecular graph* $M$ by representing its atoms as points and the covalent bonds between them as lines. Chemical graph theory ordinarily uses *hydrogen-suppressed graphs* [Trinajstić 1992], which do not include any hydrogen atoms or the bonds to them. Multiple bonds are represented by multiple lines and lone pairs of electrons by loops. (A *loop* is a line that joins a point to itself.) Different atoms can be indicated by *coloring* the points, e.g., black for carbon (●) and white for oxygen (○). A synthesis plan can be represented by a *synthesis graph* [Hendrickson 1977], in which the points stand for molecules and the lines for reactions converting one molecule into another.

We have introduced two methods to measure the complexity of (molecular) graphs. The first is the 'all possible subgraphs method,' where $N_S$ is the number of kinds of subgraphs, i.e., the number of non-isomorphic ones, and $N_T$ is the total number of subgraphs, isomorphic and non-isomorphic [Bertz & Sommer 1997, Bertz & Herndon 1986]. Only connected subgraphs are considered. The second is the 'edge cover method' [Bertz 2001, Bertz & Zamfirescu 2000], which is not discussed here. Prior to our work on the complexity of graphs, seminal contributions to this area were made by Gordon and Kennedy [1973], Minoli [1975], and Bonchev and Trinajstić [1977]. Rücker and Rücker [2001a,b] have contemporaneously made important contributions based on walk counts.

3

## 2. Results and Discussion

The problem with the methods alluded to above is their vulnerability to the 'combinatorial explosion,' e.g., the number of subgraphs increases exponentially with the number of lines. In order to simplify the problem, we have investigated subsets of all possible subgraphs such as the number of kinds of trees $T_S$ and the total number of trees $T_T$ [Bertz 2003c, Bertz & Wright 1998]. The number of spanning trees has been proposed as a measure of complexity [Gutman 1983]; however, it is not sensitive to branching [Nikolić 2003], an important complexity factor for some classes of problems. A particularly simple tree is the path on three points, $P_3$, which greatly reduces computational complexity. Gordon and Kennedy [1973] used the number of subgraphs isomorphic to $P_3$ as an index of branching, which is an essential aspect of molecular complexity. In chemical terms it is the number of ways to 'cut' the propane skeleton out of a molecule or the number of paths of length 2 in the molecular graph.

In order to extend this approach to digraphs, we must consider the possible kinds of 'paths of length 2.' There are two types of directed paths of length 2, viz. those in digraphs **A** and **D** (Figure 1). In fact, **D** contains two such paths, $p_1p_2p_1$ and $p_2p_1p_2$. In addition there are two pseudopaths, **B** and **C**. The center point in **A** has $in(p_2) = out(p_2) = 1$ and is a *carrier*, as are both points in **D**. A point is a *sink* when all arcs are directed towards it, e.g., the center point in **B**, where $out(p_2) = 0$. A point is a *source* when all arcs are directed away from it, e.g., the center point in **C**, where $in(p_2) = 0$. On the other hand, endpoints $p_1$ and $p_3$ in **B** are sources, and in **C** they are sinks.

It appears that the number of semipaths of length 2 is the simplest index of complexity that responds in a positive way to all the factors that increase the complexity of a (multi)digraph: the number of points $n$, the number of arcs $m$, the number of multiple arcs $m_k$, the number of rings (cycles) $r$, and the degree of branching $d_i$ at point $p_i$. For all trees on $n$ points, $\sum d_i = 2m$ is constant, as $m = n - 1$. Therefore, to a first approximation branching is determined by the highest degree in a graph or digraph. In cases where the highest degree is the same, the second highest may be determining. The *cyclicity* $C$ is the ratio of the number of rings $r$ to the number of points $n$, $C = r / n$. Various indices weight the complexity factors differently, and it is useful to have a range of them when confronting practical problems in order to be able to choose the index that gives the best fit.

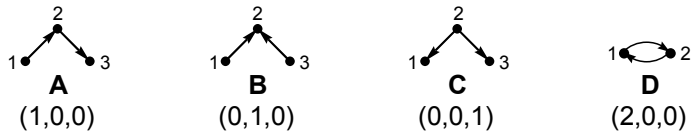


Figure 1. Possible connected digraphs with two arcs, and their triples ($a$, $b$, $c$).

Complexity and Diversity of Digraphs

For simple examples such as the digraphs in Figure 2 (vide infra), one can easily count each of the subgraphs **A**-**D** (Figure 1) and calculate the triple ($a$, $b$, $c$), where $a$ is the number of subgraphs isomorphic to **A** plus 2× the number isomorphic to **D**, $b$ is the number isomorphic to **B**, and $c$ is the number isomorphic to **C**. Below we give methods for computing ($a$, $b$, $c$), based on the adjacency matrix.

The *adjacency matrix* $A$ of a digraph $D$ on $n$ points with point set $V(D)$ and arc set $E(D)$ is the $n \times n$ matrix $A(V, E)$, where element $A(i, j) = 1$ if arc $(i, j) \in E(D)$ and $A(i, j) = 0$ otherwise. The transpose $A^{\mathrm{T}}$ of matrix $A$ is obtained by interchanging its rows and columns. The entries of $A^k(i, j)$ are the number of directed paths from $p_i$ to $p_j$ that contain $k$ arcs. Thus, we can infer that $A^2(i, j)$ contains the number of directed paths of length 2 from $p_i$ to $p_j$. Since $i = j$ is permitted, the paths in **D** are counted as well as those in **A**. Furthermore, $AA^{\mathrm{T}}(i, j)$ represents the number of pseudopaths isomorphic to **B**, where $p_i$ and $p_j$ are sources, and $A^{\mathrm{T}}A(i, j)$ represents the number of pseudopaths isomorphic to **C**, where they are sinks. Then, for digraph $D$ we obtain the triple ($a$, $b$, $c$) by computing $a$ as the sum of all entries in $A^2$, $b$ as the sum of all entries in $AA^{\mathrm{T}}$, and $c$ as the sum of all entries in $A^{\mathrm{T}}A$.

Alternatively, to reduce the computational time imposed by matrix (i.e., non-Boolean) multiplication, we note that 'local' information is sufficient to compute ($a$, $b$, $c$). The in-degree and out-degree information for every point $p$ of digraph $D$ with point set $V(D)$ can be extracted efficiently from its adjacency matrix or equivalently from an adjacency list [Gibbons 1991]. Then, $a = \sum_{p \in V} (in(p) \times out(p))$, $b = \sum_{p \in V} \mathrm{C}(in(p), 2)$ and $c = \sum_{p \in V} \mathrm{C}(out(p), 2)$, where $\mathrm{C}(k, 2)$ represents combinations of $k$ objects taken two at a time ($k$ choose 2) and the order of objects does not matter.

We have enumerated the 199 connected digraphs on four points with from zero to six 2-cycles, and they are collected in Appendices 1-5. If we neglect the directions of the arcs, then there are 53 non-isomorphic graphs underlying them. We refer to the digraphs with the same underlying graph as a *family* of digraphs. Thus, we have 53 such families with from one to sixteen members, and Figure 2 shows one example from each family (same numbering as in the appendices). Figure 3 shows all ten members of one of the families.

For complexity considerations it is useful to compute the invariant $h = a + b + c$, the total number of semipaths (i.e., paths and pseudopaths) of length 2, which is also given by equation 1, where $h_i$ is the contribution of point $p_i$ to $h$. The directionality of the arcs in a digraph is lost upon computing $h$, which is the same for the underlying graph. Therefore, $h$ characterizes the entire family of digraphs with a given underlying graph and is, in fact, precisely the Gordon-Kennedy [1973] index.

$$h = \sum_i h_i = \tfrac{1}{2} \sum_i d_i (d_i - 1) \qquad (1)$$

5

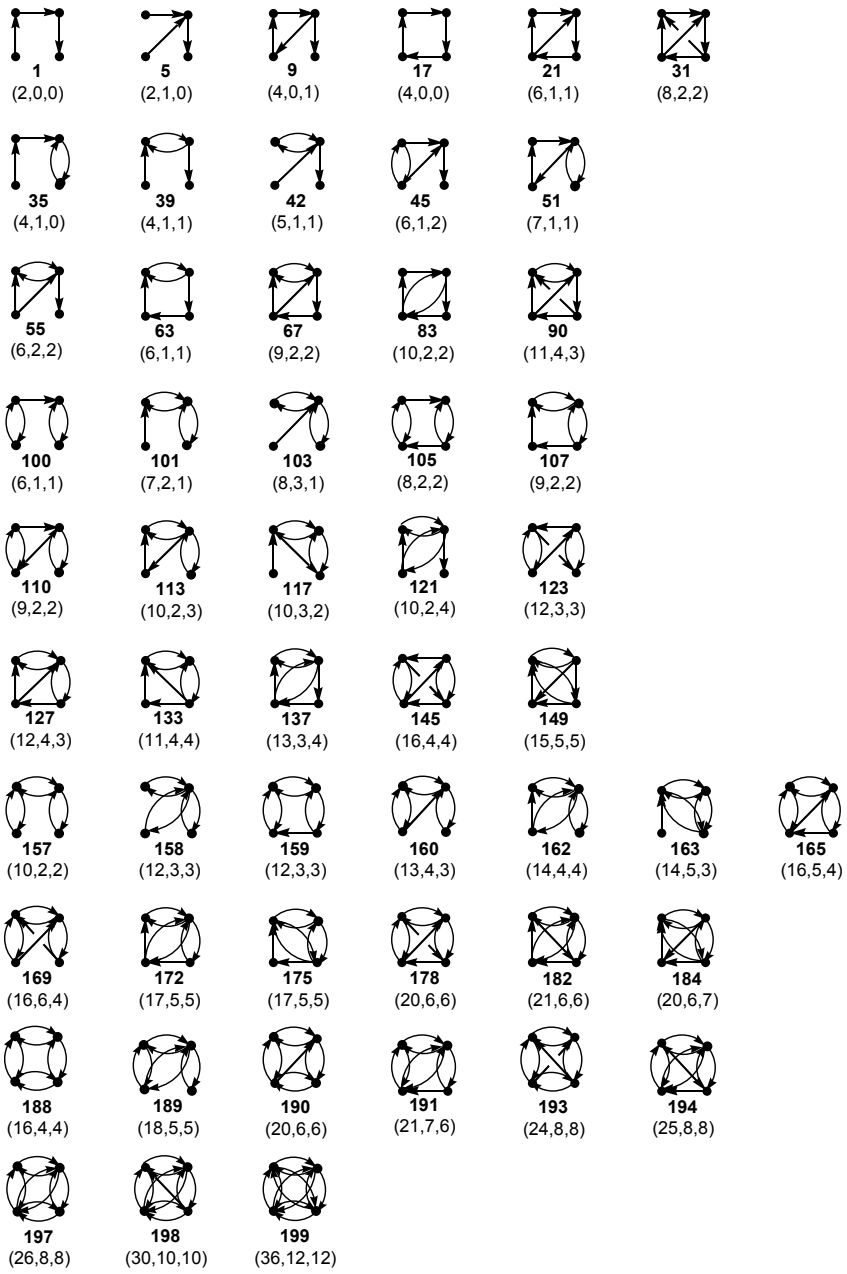


Figure 2. One example from each family of digraphs.

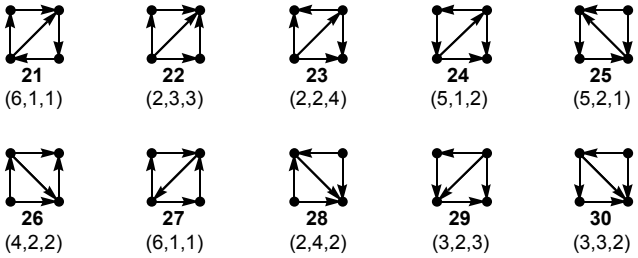


Figure 3. Digraphs with four points and five arcs (no 2-cycles).

Examining the six digraphs in the first row of Figure 2, there is a general increase in complexity as one goes from left to right, with one exception. According to $h$ values, **17** is out of place and should appear before **9**. This makes sense when one considers the complexity factors: the numbers of points ($n$ = 4), arcs ($m$ = 5) and rings ($r$ = 1) are the same, but **9** has a point of degree 3, whereas the maximum degree in **17** is 2. Cyclicity increases in the series **1**, **17**, **21** and **31** as arcs are added. Obviously, $C = r = 0$ for **1** and $C = 0.25$ ($r$ = 1) for **17**. In **21** there is one 3-cycle and one 4-cycle, and $C = 0.5$ ($r$ = 2). There are two 3-cycles and one 4-cycle in **31**, so that $C = 0.75$ ($r$ = 3). Complexity index $h$ increases in this series, and $a$, $b$ and $c$ increase monotonically.

Representative digraphs with one 2-cycle are collected in rows 2 and 3 of Figure 2, those with two of them in rows 4-6, three of them in rows 7 and 8, four in row 9, and five or six in row 10. Complexity index $h$ increases monotonically within each of these groupings except the first one, where **63** belongs between **42** and **45**. There are many examples of digraphs with the same value of $h$. There are 13 degenerate pairs, (**9**, **35**), (**45**, **51**), (**55**, **101**), (**83**, **157**), (**113**, **117**), (**127**, **133**), (**137**, **160**), (**145**, **188**), (**149**, **165**), (**162**, **163**), (**172**, **175**), (**178**, **190**) and (**182**, **184**), three triplets, (**21**, **63**, **100**), (**31**, **103**, **105**) and (**67**, **107**, **110**), and one quadruplet, (**90**, **123**, **158**, **159**). The four digraphs in the quadruplet are differentiated by the total number of trees, $T_T$ = 51, 46, 30 and 43, respectively (Appendix 6), but not by the number of kinds of trees, $T_S$ = 12, 11, 9 and 9, respectively, including $T_1 \cong P_1$. According to $T_T$, the order of decreasing complexity is **90** > **123** > **159** > **158**.

The triples ($a$, $b$, $c$) provide an indication of the diversity within a family of digraphs. Except for one pair, (**21**, **27**), all the digraphs in Figure 3 are uniquely characterized by their triples. This is the most diverse such family among those of its size or larger. For this application the order of numbers matters, e.g., (5,1,2) is not the same as (5,2,1).

The triple ($a$, $b$, $c$) also reflects the diversity of connectivity within the corresponding digraph. There are four classes of digraphs in Figure 3: **22** (2,3,3), **29** (3,2,3) and **30** (3,3,2) are more diverse than **23** (2,2,4), **26** (4,2,2) and **28** (2,4,2), which are more diverse than **24** (5,1,2) and **25** (5,2,1); they in turn are more diverse than **21** (6,1,1) and **27** (6,1,1).

7

This order is confirmed by using Shannon's 'information entropy,' $H$ (equation 2) [Shannon & Weaver 1949], where $P_i$ is the probability of semipath $i$, in this case $P_a = a / h$, $P_b = b / h$ and $P_c = c / h$. The values of $H$ for the four classes are 1.56, 1.50, 1.30 and 1.06, respectively. For this application the order of numbers in a triple does not matter.

$$H = -\sum_i P_i \log_2 P_i \qquad (2)$$

Which measure of complexity or diversity is most useful depends upon the specific application. We are especially interested in molecular complexity and synthetic complexity, i.e., the complexities of molecular graphs and synthesis digraphs, respectively. The bonds in molecules are polarized according to the electronegativities of the atoms involved, which can be modeled by using arcs in the *polarity digraph*, where an arc goes from the less to the more electronegative atom. Electronegativity is one of the main factors governing chemical reactivity, e.g., it predicts the site of attack by hydroxide on an ester, as illustrated in Figure 4. (Strictly speaking, **1** and **3** are multidigraphs, vide supra.)

Quantities $a$, $b$ and $c$ are not equally relevant to all problems. A *synthesis digraph* represents a multistep synthetic plan. For this application, only directed paths from the available starting materials to the desired product are fruitful [Bertz & Sommer 1993, Bertz 1986], and $a$ is the most important parameter. As far as $h$ is concerned, a similar quantity, η, the number of pairs of adjacent lines [Bertz 1981a,b], has been used as a predictor of relative synthetic efficiency [Bertz & Wright 1998, Bertz 1982, 1983].

## 3. Conclusion

We have demonstrated that the simplest approach to the complexity of graphs, the number of paths of length 2, gives useful results when extended to the complexity and diversity of digraphs. As expected for such a primitive method, the discriminating power is not as great as more sophisticated approaches. For some applications, the individual quantities $a$, $b$ and $c$ may be more useful than their sum, $h$. All four invariants are involved in calculating the information entropy, $H$, which is a useful measure of diversity for digraphs.

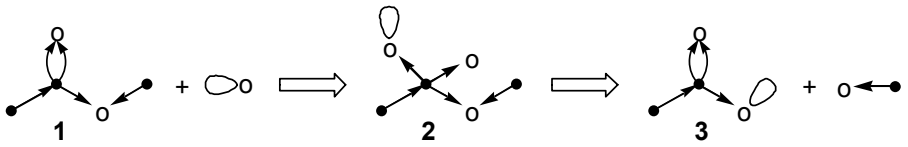


Figure 4. Polarity digraphs for the alkaline hydrolysis ($^{-}$OH) of an ester, methyl acetate. Negative charges are not shown, but are associated with the lone pairs of electrons (loops).

## Appendices

Appendices 1-6 are available from the first author upon request.

## References

Bar-Yam, Y. (Ed.), 2000, *Unifying Themes in Complex Systems*, Westview (Boulder, CO).
Bar-Yam, Y., & Minai, A. (Eds.), 2003, *Unifying Themes in Complex Systems, vol. 2*, Westview (Boulder, CO).
Bertz, S.H., 2003a, *New J. Chem*., **27**, 860-869.
Bertz, S.H., 2003b, *New J. Chem*., **27**, 870-879.
Bertz, S.H., 2003c, *Complexity in Chemistry: Introduction and Fundamentals*, Bonchev, D., & Rouvray, D.H. (Eds.), Taylor & Francis (London) 91-156.
Bertz, S.H., 2001, *Chem. Commun*., 2516-2517.
Bertz, S.H., 1986, *J. Chem. Soc., Chem. Commun*., 1627-1628.
Bertz, S.H., 1983, *Chemical Applications of Topology and Graph Theory*, King, R.B. (Ed.), Elsevier (Amsterdam) 206-221.
Bertz, S.H., 1982, *J. Am. Chem. Soc*., **104**, 5801-5803.
Bertz, S.H., 1981a, *J. Am. Chem. Soc*., **103**, 3599-3601.
Bertz, S.H., 1981b, *J. Chem. Soc., Chem. Commun*., 818-820.
Bertz, S.H., & Herndon, W.C., 1986, *Artificial Intelligence Applications in Chemistry*, Pierce, T.H., & Hohne, B.A. (Eds.), American Chemical Society (Washington, DC) 169-175.
Bertz, S.H., & Sommer, T.J., 1997, *Chem. Commun*., 2409-2410.
Bertz, S.H., & Sommer, T.J., 1993, *Organic Synthesis: Theory and Applications, vol. 2*, Hudlicky, T. (Ed.), JAI Press (Greenwich, CT) 67-92.
Bertz, S.H., & Wright, W.F., 1998, *Graph Theory Notes of New York (NY Acad. Sci.)*, **XXXV**, 32-48.
Bertz, S.H., & Zamfirescu, C.M., 2000, *MATCH–Commun. Math. Comput. Chem.*, **42**, 39-70.
Bonchev, D., & Trinajstić, N., 1977, *J. Chem. Phys*., **67**, 4517-4533.
Cowan, G.A., Pines, D., & Meltzer D. (Eds.), 1994, *Complexity, Models, and Reality*, Addison-Wesley (Reading, MA).
Gibbons, A., 1991, *Algorithmic Graph Theory*, Cambridge University Press (Cambridge, UK).
Gordon, M., & Kennedy, J.W., 1973, *J. Chem. Soc., Faraday Trans. II*, **69**, 484-504.
Gutman, I., Mallion, R.B., & Essam, J.W., 1983, *Mol. Phys*., **50**, 859-877.
Harary, F., 1969, *Graph Theory*, Addison-Wesley (Reading, MA).
Hendrickson, J.B., 1977, *J. Am. Chem. Soc*., **99**, 5439-5450.
Minoli, D., 1975, *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (Ser. 8)*, **59**, 651-661.
Nadel, L., & Stein, D.L. (Eds.), 1995, *1993 Lectures in Complex Systems*, Addison-Wesley (Redwood City, CA).
Nikolić, S., Trinajstić, N., Tolić, I.M., Rücker, G., & Rücker, C., 2003, *Complexity in Chemistry: Introduction and Fundamentals*, Bonchev, D., & Rouvray, D.H. (Eds.), Taylor & Francis (London) 29-89.
Rücker, G., & Rücker, C., 2001a, *J. Chem. Inf. Comput. Sci*., **41**, 314-320.
Rücker, G., & Rücker, C., 2001b, *J. Chem. Inf. Comput. Sci*., **41**, 1457-1462.
Shannon, C.A., & Weaver, W., 1949, *The Mathematical Theory of Communication*, University of Illinois Press (Urbana).
Trinajstić, N., 1992, *Chemical Graph Theory, 2nd edn*., CRC Press (Boca Raton, FL).