communications in mathematical
match
and in computer chemistry

no. 42, October 2000

ISSN 0340-6253

MATCDY (42) 39-70 (2000)

New Complexity Indices based on Edge Covers

Steven H. Bertz* and Christina M. Zamfirescu†

*Complexity Study Center
88 East Main Street, Box 220
Mendham, NJ 07945

†Hunter College, CUNY
Computer Science Department
695 Park Avenue
New York, NY 10021

(Received June 1999)

Abstract

After abstracting a system as a graph G, its complexity can be measured by using graph invariants $I(G)$, which represent the interactions within the system. An index of intrinsic complexity is characterized as an invariant that increases in homologous series such as paths $\{P_n\}$, cycles $\{C_n\}$, star graphs $\{K_{1,n}\}$, and complete graphs $\{K_n\}$. Furthermore, it obeys the inequalities $I(K_n) \geq I(C_n) \geq I(P_n)$ and $I(K_n) \geq I(K_{1,n-1}) \geq I(P_n)$ for isomeric graphs on n vertices. Extrinsic complexity indices, which are usually based on information about how the system is constructed, often do not satisfy these conditions. New complexity indices based on edge covers and partitions are introduced, e.g., k_S^{bi}, the number of kinds of minimal edge biclique covers, k_T^{bi}, the total number of minimal edge biclique covers, p_T^{bi}, the total number of partitions of G into bicliques, and $\bar{I}[L^2(G)]$, the intersection number of the second line graph. Graph invariants are used to induce partial orders on selected sets of graphs, and the resulting posets are studied with the aid of Hasse diagrams.

1. Introduction

There has been a great deal of interest in complexity in recent years [1-13]. Physics, chemistry and biology are increasingly being viewed in terms of hierarchies of complexity characterized by emergent phenomena and broken symmetry [10-13]. The two main approaches to measuring the complexity of an object are (i) to examine the way it is constructed, either actually or conceptually, and (ii) to study the way its components interact. The theme of this paper is approach (ii), which we call *intrinsic complexity*. We use *extrinsic complexity* to describe approach (i). With either approach a partial order [14-16] may result when two or more objects are compared. We introduce several new intrinsic complexity indices and discuss some previous indices of both kinds. Our method is to abstract the system under study—e.g., a molecule, a synthesis plan for its preparation, or a map of its rearrangements—as a graph and then use the tools of graph theory [17,18] and combinatorics [16,19] to understand the relationships in it.

2. Definitions

2.1 Mathematical concepts

The definitions given below follow standard texts [15-19]. 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 $V(G)$. If both $V(G)$ and $E(G)$ are the null set ϕ, the result is the *null graph* G_0. Each edge $x = uv = vu$ in G is said to *join* u and v. We say that u and v are *adjacent* vertices. Two edges are *adjacent* when they share a vertex, e.g., uv and vw. Vertex u and edge x are said to be *incident* to each other. (Many authors also use *adjacent* here.) The *degree* d of a vertex is the number of edges incident to it. In a *regular graph* all the vertices have the same degree. We use vertex and point interchangeably and also edge and line, but try not to mix the two systems.

There are several variations on the theme of graphs relevant here. If the edges are directed *arcs*, i.e., ordered pairs of distinct vertices, it is a *directed graph* or *digraph D*.

An (undirected) edge in graph G can be considered to be a superposition of two arcs pointing in opposite directions. In a *multigraph* more than one edge (i.e., a *multiple edge*) joins at least one pair of vertices. A *loop* $x = uu$ is an edge joining a vertex to itself. When constructing line graphs (vide infra), we do not consider a loop to be adjacent to itself. A *pseudograph* allows both multiple edges and loops.

A *cycle* C_n is a sequence of vertices $v_1, v_2, v_3, \ldots, v_n$, which are joined by edges v_1v_2, $v_2v_3, \ldots, v_{n-1}v_n$ such that the first vertex v_1 and last vertex v_n are joined by edge v_1v_n. Otherwise, it is a *path* on n vertices, P_n. In this paper paths are *simple* or *self-avoiding*, i.e., no vertex appears more than once in the sequence of vertices. In a *connected* (1-component) graph all pairs of vertices are the endpoints of some path.

A graph G is *labeled* when its n points are distinguished from each other by labels, say $\lambda_1, \lambda_2, \ldots, \lambda_n$. Each point has a unique label, e.g., used for isomorphism testing (vide infra). Another common way to label graphs allows more than one point to have the same label. To avoid confusion these non-unique labels are called *colors*, and to assign them is *to color* the points. A *coloring* of a graph assigns colors to its points in such a way that no pair of adjacent points has the same color. If n colors are used, an *n-coloring* results and partitions $V(G)$ into n *color classes*, which are the sets of (independent) points with the same color.

A *bipartite graph* or *bigraph* has a 2-coloring in which every line connects a point in one color class to a point in the other. A *tree* is a connected graph without cycles. All trees are bipartite.

A *subgraph* S of G is a graph that has all its vertices in $V(G)$ and edges in $E(G)$. We include G itself in the set of all its possible subgraphs. A *spanning subgraph* of G is a subgraph containing all the vertices of G. A spanning subgraph that is also a tree is a *spanning tree*. The removal of an edge x from G yields the spanning subgraph $G - x$ containing all edges of G except x. For any subset S of $V(G)$, the *induced subgraph* $\langle S \rangle$ or *subgraph of G induced by S* is the maximal subgraph of G with vertex set S, i.e., two vertices of S are adjacent in $\langle S \rangle$ iff they are adjacent in G.

If G_1 is a subgraph of G_2, then G_2 is a *supergraph* of G_1. If uv is not an edge in G_1, but u is a vertex in this graph, then the addition of edge $x = uv$ to G_1 results in the smallest supergraph G_2 of G_1 containing the edge uv. We write $G_2 = G_1 + x$. When constructing homologous series, we often add a vertex v in this way (section 3.3).

Two graphs G and H are *isomorphic*, $G \cong H$, iff there exists a one-to-one correspondence between their point sets that preserves adjacency. Isomorphism testing requires labeling the points (vide supra). An *invariant* of G is a number $I(G)$ associated with G that has the same value for any graph H isomorphic to G. As a chemical example, two conformations of a molecule have isomorphic molecular graphs, e.g., chair and boat cyclohexane (C_6) or trans and gauche butane (P_4).

Molecules are *isomers* whenever they have the same number and kinds of atoms (i.e., the same molecular formula) and the same number of bonds. By analogy, we call two vertex colored graphs *isomeric* iff they have the same number of vertices of each color and the same number of edges. Examples of isomeric graphs (one color) include P_4 and $K_{1,3}$, C_3 and $C_2 + x$, and C_4 and $C_3 + x$.

The *line graph* $L(G)$ of a graph (or pseudograph) G has the edges of G as its vertices, and two vertices in $L(G)$ are adjacent whenever the corresponding edges in G are adjacent. Thus, the pairs of adjacent lines in G become the lines in $L(G)$. This concept is so natural that it has been discovered or invented many times and been given many names [17]: the derived graph, edge-to-vertex dual, covering graph, adjoint, graph derivative, interchange graph and, for chemical applications, the bond graph. Repeating the process embodied in the definition leads to the *iterated line graphs* $L^n(G)$, where $L^0(G) \cong G$ and $L^1(G) \cong L(G)$. We consider $L(C_1) \cong P_1$, $L(P_1) \cong P_0$, and $L(P_0)$ is not defined. We refer to $L(G)$ and $L^2(G)$ as the first and second line graph, respectively. (N.B., Beineke [20] used $\partial(D)$ and $\partial^2(D)$, where D is a digraph, and Menon [21] used $I(G)$ and $I^2(G)$, respectively.) Pictures of a number of the line graphs discussed herein have been published [22]. The introduction of the line graph enables multigraphs and pseudographs to be treated by using indices defined for graphs (cf. section 2.2).

A *complete graph* K_n has every pair of its n points adjacent, i.e., each point is joined to every other point by one of $n(n-1)/2$ lines. Following reference [18], we allow the null graph $K_0 \cong K_{0,0} \cong P_0$, which has $V(K_0) = \phi$ and $E(K_0) = \phi$. However, we do not ordinarily include K_0 in the set of all possible subgraphs. A *clique* in G is a complete subgraph of G. A *complete bipartite graph* or *complete bigraph* $K_{m,n}$ is a bigraph that contains m points of one color, n points of the other, and all the possible mn lines. A complete bipartite subgraph of G is a *biclique* in G. Examples of common bicliques in molecular graphs include $K_{2,2} \cong C_4$ (cyclobutane) and the star graphs $K_{1,n}$.

An *edge cover* of G is any family $\mathfrak{C} = \{S_1, \ldots, S_n\}$ of subgraphs S_i of G such that every edge of G is in at least one of the $E(S_i)$, $i = 1, \ldots, n$. In an *edge clique cover* each $S_i \in \mathfrak{C}$ is a clique in G; in an *edge biclique cover* each $S_i \in \mathfrak{C}$ is a biclique in G. A *set system* is an edge cover where no $S_i \subseteq S_j$ for all $i, j = 1, \ldots, n$ $(i \neq j)$. A *minimal edge cover* does not properly contain any other edge cover. Thus, in a minimal edge cover every $S_i \in \mathfrak{C}$ is essential, i.e., S_i covers at least one edge in $E(G)$ that is not covered by any other subgraph $S_j \in \mathfrak{C}$ $(i \neq j)$. As a test, the exclusion of one S_i at a time results at each step in a family of subgraphs that no longer covers $E(G)$. A *partition* P is an edge cover with the additional property that every edge belongs to exactly one $S_i \in \mathfrak{C}$.

A set $\hat{\mathrm{G}}$ of graphs has a *composition series* if there exists a countable sequence $<G_1, G_2, G_3, \ldots>$ of graphs in $\hat{\mathrm{G}}$ such that each G_i is an induced subgraph of G_{i+1} and every $G \in \hat{\mathrm{G}}$ is the induced subgraph of some G_i. A *homologous series* $\{H\}$ is generated by incrementing a connected graph according to a *recurrent rule*. When a new vertex v is added, one or more new edges are also added to join v to the other vertices as specified by the rule. The sets of all paths $\{P_n\}$, star graphs $\{K_{1,n}\}$, and complete graphs $\{K_n\}$ have composition series and are homologous series. The set of all cycles $\{C_n\}$ is a homologous series, but does not have a composition series (of connected graphs). The set of all planar graphs has a composition series, but is not a homologous series.

The ordered pair (S, $\mathfrak{R}$) is a *partially ordered set* or *poset* when S is a set and $\mathfrak{R}$ is a relation on S that is reflexive, antisymmetric and transitive. For our purposes the relation will be the familiar $\geq$ (greater than or equal to). For all X, Y, $Z \in S$ the reflexive property is $X \geq X$ (i.e., $X = X$); the antisymmetric property stipulates that if $X \geq Y$ and $Y \geq X$, then $X = Y$; and the transitive property requires that if $X \geq Y$ and $Y \geq Z$, then $X \geq Z$. If the elements of every pair X, $Y \in S$ can be ordered by $\mathfrak{R}$, then (S, $\mathfrak{R}$) is a *total order*.

A *Hasse diagram* of the poset (S, $\geq$) is a digraph D with $S = V(D)$ as its vertex set (s, t, $u \in S$, $s \neq t$, $s \neq u$, $t \neq u$) and with a directed edge from s to t iff $s \geq t$, and also there is no u in S with $s \geq u \geq t$. The directions of the edges are customarily omitted as unnecessary, since Hasse diagrams are drawn with their edges in conventional directions.

A molecule can be abstracted as a *molecular graph* M by representing atoms as vertices and covalent bonds between atoms as edges. Chemical graph theory ordinarily uses *skeletal* or *hydrogen-suppressed* molecular graphs, which do not include any hydrogen atoms. They resemble the abbreviated structure drawings commonly used in organic chemistry. Multiple bonds are represented by multiple edges and unshared pairs of electrons by loops. Heteroatoms can be included by coloring the vertices, where each chemical element is assigned a unique color. A synthesis plan can be represented by a *synthesis digraph*, in which the points stand for molecules and the arcs for reactions that convert one molecule into another [3,23]. Since the arcs always point in the forward (synthetic) direction, they are usually drawn without the arrowheads in a *synthesis graph*, which is important for its symmetry properties [3,24].

2.2 Topological indices

A *topological index* is defined as a number (integer or real) that is a graph invariant or is derived from one or more invariants. It does not change if the graph is replaced by an isomorphic one, as happens when it is relabeled, redrawn or changed in any way that leaves the adjacency relation intact. There are many papers on topological indices; we only give leading references [25-31].

The *intersection number* $i(G)$ of a graph G equals the minimum number of cliques needed to cover its edges. (N.B., this is not its original definition, which was based on intersection graphs [18], but an operational one that provides a way to count it [32].) We define the *bintersection number* $j(G)$ as the minimum number of bicliques needed to cover the edges of a graph G. The analogy between $i(G)$ and $j(G)$ first arose in computing the intersection numbers of digraphs [33-35]. Table 1 summarizes $i[L^n(G)]$ and $j[L^n(G)]$, $n = 0, 1, 2$, for important homologous series.

We introduce the number of kinds of minimal edge clique covers k_S, the total number of minimal edge clique covers k_T, and the bipartite analogs, the number of kinds of minimal edge biclique covers k_S^{bi} and their total number k_T^{bi}. Two covers are of the same kind when they consist of the same cliques (or bicliques) in equal numbers. In contrast to subgraphs (next paragraph) non-isomorphic covers can be of the same kind; e.g., $K_{1,3} + K_{1,2} + K_{1,1}$ comprises two non-isomorphic minimal edge biclique covers of $K_4 - x$ (see unlabeled graphs, Table 4). For the total number, a graph is considered to be labeled (cf. section 3.6). The numbers of kinds of partitions p_S and p_S^{bi} and the total numbers of partitions p_T and p_T^{bi} for cliques and bicliques, respectively, are also considered. They are compiled in Table 2, except for p_S and p_T (see section 3.5).

Some previous topological indices are collected in Table 3. The Tutte (N^T_{span}), Minoli (χ_M) and Mowshowitz (H) approaches are discussed in section 3.1. The number of kinds of (connected) subgraphs N_S counts distinct, non-isomorphic subgraphs, and the total number of (connected) subgraphs N_T counts all possible ones [2,36]. (Note the difference between kinds of subgraphs and kinds of covers.) Every point P_1 is considered to be a subgraph, as is the graph itself. Defined in this way, N_S and N_T can be calculated from concise formulas for specific homologous series (section 3.3). The number of ways to 'cut' propane out of a molecule ($N_{2,j}$) was an early branching index [26]; it is one-half Platt's f index [27]. This invariant has been generalized to include multiple edges with the aid of line graphs: the *connections* are the pairs of adjacent edges in G or the edges in $L(G)$ [37,38]. This identity is predicated on the definition of the line graph of a multigraph that counts a double edge as one pair of adjacent edges [37]. (An alternative definition counts a double edge as two adjacencies [20,21].)

Equation 2.2.1 for $C(n)$ is an extension of the 'information entropy' approach [39,40] (cf. section 3.1), where n_i is the cardinality of F_i, the ith family of equivalent subgraphs, and n is the total number of subgraphs. When the subgraphs are edges in $L(G)$, $n = \eta$ (connections), $n_i = \eta_i$ (equivalent connections), and $C(\eta)$ is calculated according to equation 2.2.2 [38]. While η is an index of intrinsic complexity, $C(\eta)$ is not, as it does not strictly follow the hierarchies of homologous series (see section 3.4).

(2.2.1) $$C(n) = 2n \log_2 n - \sum_i n_i \log_2 n_i$$

(2.2.2) $$C(\eta) = 2\eta \log_2 \eta - \sum_i \eta_i \log_2 \eta_i$$

2.3 Sample calculations for topological indices

We derive the above indices for C_4, which represents the molecule cyclobutane. The largest clique is K_2, and four of these are required for the only minimal edge clique cover, $4K_2$. Furthermore, $4K_2$ is also a partition; thus, $k_S = p_S = 1$ and $k_T = p_T = 1$. (Other edge covers are $4K_2 + K_1$, $4K_2 + 2K_1$, $4K_2 + 3K_1$ and $4K_2 + 4K_1$; however, they are not minimal.) By default, $4K_2$ is the smallest edge clique cover, and $i(C_4) = 4$. Since $L(C_4) \cong C_4$, $i[L^n(C_4)] = 4$ for all n.

The largest biclique in C_4 is $K_{2,2} \cong C_4$, and it is a minimal edge biclique cover. There are two other non-trivial bicliques in C_4 ($K_{1,2}$ and $K_{1,1}$), and the five kinds of minimal edge biclique covers (number of each kind) are $K_{2,2}$ (1), $2K_{1,2}$ (2), $2K_{1,2} + K_{1,1}$ (4), $K_{1,2} + 2K_{1,1}$ (4), and $4K_{1,1}$ (1). Therefore, $k_S^{bi} = 5$ and $k_T^{bi} = 12$, which is the sum of the numbers in parentheses. All these covers are partitions, except $2K_{1,2} + K_{1,1}$; therefore, $p_S^{bi} = 4$ and $p_T^{bi} = 8$. (If we were to include set systems, $4K_{1,2}$ (1) and $3K_{1,2}$ (4) would be added; see also section 3.6.) Since $L(C_4) \cong C_4 \cong K_{2,2}$, the minimum number of bicliques needed to cover the edges of C_4 is $j(C_4) = 1$. Indeed, $j[L^n(C_4)] = 1$ for all n.

Previous indices for C_4 are given for comparison. There are four (equivalent) connections; therefore, $\eta = 4$ and $C(\eta) = 2 \times 4 \log_2 4 - 4 \log_2 4 = 8$. The kinds of

subgraphs (number of each kind) are P_1 (4), P_2 (4), P_3 (4), P_4 (4) and C_4 (1); thus, N_S = 5 and N_T = 17. (Note the origin of the $N_S = n + 1$ and $N_T = n^2 + 1$ formulas for C_n; see also section 3.3.) Finally, $N^T_{span} = 4$, $\chi_M = 24.00$ and $H = 0$ (cf. section 3.1).

A more complicated example is $K_4 - x$, which represents the molecule bicyclobutane. There are three kinds of minimal edge clique covers (number of each kind): $2K_3$ (1), $K_3 + 2K_2$ (2) and $5K_2$ (1). Thus, $k_S = 3$ and $k_T = 4$. Furthermore, the latter two covers are partitions, so that $p_S = 2$ and $p_T = 3$. The minimal edge biclique covers are summarized in Table 4, from which we derive $k_S^{bi} = 15$, $p_S^{bi} = 6$, $k_T^{bi} = 90$ and $p_T^{bi} = 24$. (For the total numbers k_T^{bi} and p_T^{bi}, the graph is considered to be labeled; see also section 3.6.)

Since $K_4 - x$ is not a complete graph (bigraph), the smallest possible number of cliques (bicliques) hypothetically able to cover its edges is two. There is such an edge clique cover, $2K_3$; consequently, $i(K_4 - x) = 2$. Also, the smallest number of bicliques needed to cover its edges is $j(K_4 - x) = 2$ (e.g., $K_{2,2} + K_{1,1}$, see Table 4). The line graph $L(K_4 - x)$ is pictured in [22]; it can be constructed by joining a new vertex v to each vertex of C_4. A smallest edge clique cover (e.g., $4K_3$ or $2K_3 + 2K_2$) has cardinality 4, which gives $i[L(K_4 - x)] = 4$. Since $K_{2,2} + K_{1,4}$ is the smallest edge biclique cover, $j[L(K_4 - x)] = 2$. The second line graph $L^2(K_4 - x)$ is straightforward to construct and is also pictured in [22]. It can be covered by a minimum of 5 cliques ($K_4 + 4K_3$) or 5 bicliques (e.g., $K_{2,2} + 4K_{1,5}$ or $K_{2,2} + 4K_{2,3}$); therefore, $i[L^2(K_4 - x)] = j[L^2(K_4 - x)] = 5$.

As far as the previous indices for $K_4 - x$ are concerned, $\eta = 8$ and $C(\eta) = 2 \times 8 \log_2 8 - 4 \log_2 4 - 2 \log_2 2 - 2 \log_2 2 = 36$. (The number 32, a typographical error in the first full paper on this index [4], was copied into a recent review [2]; it is corrected here.) The kinds of subgraphs (number of each kind) are P_1 (4), P_2 (5), P_3 (8), C_3 (2), P_4 (6), C_4 (1), $K_{1,3}$ (2), $C_3 + x$ (4), and $K_4 - x$ (1); therefore, $N_S = 9$ and $N_T = 33$. (By $C_3 + x$ we mean the graph of methylcyclopropane, the smallest supergraph of C_3; see section 2.1.) Finally, $N^T_{span} = 8$, $\chi_M = 42.22$ and $H = 1.00$ (cf. section 3.1).

All the indices, new and old, discussed above give the correct order of complexity, viz. bicyclobutane ($K_4 - x$) > cyclobutane (C_4) (see section 3.2), except for $i(G)$ and $i[L(G)]$.

Table 1. Intersection and bintersection numbers for some homologous series.[a]

G	$i(G)$	$i[L(G)]$	$i[L^2(G)]$	$j(G)$	$j[L(G)]$	$j[L^2(G)]$
P_1	1	1[b]	—[c]	1	1[b]	—[c]
P_2	1	1	1[b]	1	1	1[b]
P_3	2	1	1	1	1	1
P_4	3	2	1	2	1	1
P_5	4	3	2	2	2	1
P_6	5	4	3	3	2	2
P_7	6	5	4	3	3	2
P_8	7	6	5	4	3	3
C_1	1	1	1[b]	1	1	1[b]
C_2	2	1	1	1	1	1
C_3	1	1	1	2	2	2
C_4	4	4	4	1	1	1
C_5	5	5	5	3	3	3
C_6	6	6	6	3	3	3
C_7	7	7	7	4	4	4
C_8	8	8	8	4	4	4
$K_{1,0}$	1	1[b]	—[c]	1	1[b]	—[c]
$K_{1,1}$	1	1	1[b]	1	1	1[b]
$K_{1,2}$	2	1	1	1	1	1
$K_{1,3}$	3	1	1	1	2	2
$K_{1,4}$	4	1	4	1	2	2
$K_{1,5}$	5	1	5[d,e]	1	3[d,f]	5[d,g]
K_1	1	1[b]	—[c]	1	1[b]	—[c]
K_2	1	1	1[b]	1	1	1[b]
K_3	1	1	1	2	2	2
K_4	1	4[d,h]	6[d,i]	2[d,j]	2[d,k]	8[d,l]

[a]Chemists N.B.: $P_1 \cong K_{1,0} \cong K_1$ = methane, $P_2 \cong K_{1,1} \cong K_2$ = ethane, $P_3 \cong K_{1,2}$ = propane, P_4 = butane, P_5 = pentane, P_6 = hexane, P_7 = heptane, P_8 = octane, C_1 = methylene (carbene), C_2 = ethylene, $C_3 \cong K_3$ = cyclopropane, C_4 = cyclobutane, C_5 = cyclopentane, C_6 = cyclohexane, C_7 = cycloheptane, C_8 = cyclooctane, $K_{1,3}$ = 2-methylpropane (isobutane), $K_{1,4}$ = 2,2-dimethylpropane, $K_{1,5}$ = pentamethylmethane (hypothetical), K_4 = tetrahedrane. [b]The index is given for $P_0 \cong K_{0,0} \cong K_0$. [c]Not defined. [d]This is the smallest number of covering cliques or bicliques we have been able to find; there may be a smaller one. [e]E.g., $5K_4$. [f]E.g., $K_{2,3} + K_{2,2} + K_{1,2}$ or $2K_{2,3} + K_{1,1}$ or $3K_{2,2}$. [g]E.g., $5K_{2,4}$. [h]E.g., $4K_3$. [i]E.g., $6K_4$. [j]E.g., $2K_{2,2}$. [k]E.g., $K_{2,4} + K_{2,2}$ or $2K_{2,4}$. [l]E.g., $4K_{1,6} + 4K_{2,2}$ or $6K_{2,3} + 2K_{1,1}$.

Table 2. Indices based on minimal edge clique and biclique covers and partitions.[a]

G	$k_S(G)$	$k_T(G)$	$k_S^{bi}(G)$	$k_T^{bi}(G)$	$p_S^{bi}(G)$	$p_T^{bi}(G)$
P_1	1	1	1	1	1	1
P_2	1	1	1	1	1	1
P_3	1	1	2	2	2	2
P_4	1	1	3	4	2	3
P_5	1	1	4	7	4	5
P_6	1	1	5	13	5	8
P_7	1	1	7	24	9	13
P_8	1	1	8	44	12	21
C_1	1	1	1	1	1	1
C_2	1	1	2	3	2	3
C_3	2	2	3	7	2	4
C_4	1	1	5	12	4	8
C_5	1	1	5	21	3	11
C_6	1	1	7	39	5	18
C_7	1	1	8	71	5	29
C_8	1	1	10	131	8	47
$K_{1,0}$	1	1	1	1	1	1
$K_{1,1}$	1	1	1	1	1	1
$K_{1,2}$	1	1	2	2	2	2
$K_{1,3}$	1	1	4	8	3	5
$K_{1,4}$	1	1	9	49	5	15
$K_{1,5}$	1	1	20	522	7	52
K_1	1	1	1	1	1	1
K_2	1	1	1	1	1	1
K_3	2	2	3	7	2	4
K_4	5	16	—[b]	—[b]	7	70

[a]Chemists N.B.: $P_1 \cong K_{1,0} \cong K_1$ = methane, $P_2 \cong K_{1,1} \cong K_2$ = ethane, $P_3 \cong K_{1,2}$ = propane, P_4 = butane, P_5 = pentane, P_6 = hexane, P_7 = heptane, P_8 = octane, C_1 = methylene (carbene), C_2 = ethylene, $C_3 \cong K_3$ = cyclopropane, C_4 = cyclobutane, C_5 = cyclopentane, C_6 = cyclohexane, C_7 = cycloheptane, C_8 = cyclooctane, $K_{1,3}$ = 2-methylpropane (isobutane), $K_{1,4}$ = 2,2-dimethylpropane, $K_{1,5}$ = pentamethylmethane (hypothetical), K_4 = tetrahedrane. [b]Not yet calculated.

Table 3. Previous indices for some homologous series.[a]

G	η	$C(\eta)$	N_S	N_T	N'_{span}	$\chi_M(G)$
P_1	0	—[b]	1	1	1	0
P_2	0	—[b]	2	3	1	0.67
P_3	1	0	3	6	1	3.60
P_4	2	2.00	4	10	1	10.29
P_5	3	7.51	5	15	1	22.22
P_6	4	12.00	6	21	1	40.91
P_7	5	19.22	7	28	1	67.85
P_8	6	25.02	8	36	1	104.53
C_1	0	—[b]	2	2	1	0
C_2	1	0	3	5	2	2.00
C_3	3	4.75	4	10	3	9.00
C_4	4	8.00	5	17	4	24.00
C_5	5	11.61	6	26	5	50.00
C_6	6	15.51	7	37	6	90.00
C_7	7	19.65	8	50	7	147.00
C_8	8	24.00	9	65	8	224.00
$K_{1,0}$	0	—[b]	1	1	1	0
$K_{1,1}$	0	—[b]	2	3	1	0.67
$K_{1,2}$	1	0	3	6	1	3.60
$K_{1,3}$	3	4.75	4	11	1	10.29
$K_{1,4}$	6	15.51	5	20	1	22.22
$K_{1,5}$	10	33.22	6	37	1	40.91
K_1	0	—[b]	1	1	1	0
K_2	0	—[b]	2	3	1	0.67
K_3	3	4.75	4	10	3	9.00
K_4	12	43.02	10	64	16	72.00

[a]Chemists N.B.: $P_1 \cong K_{1,0} \cong K_1$ = methane, $P_2 \cong K_{1,1} \cong K_2$ = ethane, $P_3 \cong K_{1,2}$ = propane, P_4 = butane, P_5 = pentane, P_6 = hexane, P_7 = heptane, P_8 = octane, C_1 = methylene (carbene), C_2 = ethylene, $C_3 \cong K_3$ = cyclopropane, C_4 = cyclobutane, C_5 = cyclopentane, C_6 = cyclohexane, C_7 = cycloheptane, C_8 = cyclooctane, $K_{1,3}$ = 2-methylpropane (isobutane), $K_{1,4}$ = 2,2-dimethylpropane, $K_{1,5}$ = pentamethylmethane (hypothetical), K_4 = tetrahedrane. [b]Not defined.

Table 4. Minimal edge biclique covers for $K_4 - x$ (bicyclobutane).

entry[a]	cover	unlabeled[b]	labeled[c]
1	$K_{2,2} + K_{1,3}$	1	2
2	$K_{2,2} + K_{1,2}$	1	4
3 (partition)	$K_{2,2} + K_{1,1}$	1	1
4	$2K_{1,3}$	1	1
5	$K_{1,3} + 2K_{1,2}$	3	8
6	$K_{1,3} + K_{1,2} + K_{1,1}$	2	8
7 (partition)	$K_{1,3} + K_{1,2}$	1	2
8 (partition)	$K_{1,3} + 2K_{1,1}$	1	2
9	$4K_{1,2}$	1	1
10	$3K_{1,2}$	5	16
11	$3K_{1,2} + K_{1,1}$	2	8
12	$2K_{1,2} + 2K_{1,1}$	6	18
13 (partition)	$2K_{1,2} + K_{1,1}$	4	10
14 (partition)	$K_{1,2} + 3K_{1,1}$	3	8
15 (partition)	$5K_{1,1}$	1	1

[a]The highest entry number is k_S^{bi}. [b]The column sum is k_U^{bi}. [c]The column sum is k_T^{bi}.

3. Complexity

3.1 Previous approaches based on graph theory

Three early approaches to the complexity of graphs were Rashevsky's 'information entropy' H [39], which was associated with complexity by Mowshowitz [40], the number of spanning trees N^T_{span}, which was used to measure the complexity of networks by Tutte [41], and the combinatorial complexity function $\chi_M(G)$ of Minoli [42]. In equation 3.1.1 for $\chi_M(G)$, n is the number of vertices, e is the number of edges, and $\sigma_{i,j}$ is the number of paths between vertices i and j. Minoli's function is constant for isomeric trees (e.g., χ_M = 10.29 for P_4 and $K_{1,3}$), but increases with the size of the tree (number of vertices, see Table 3). Trees have but one spanning subgraph, the tree itself; hence, N^T_{span} = 1 for all trees. The 'information entropy' H is based on Shannon's equation 3.1.2, applied to families F_i of equivalent points [39] (orbits of the automorphism group [40]), where $p_i = n_i/n$, n_i is the cardinality of F_i, and n is the total number of points. This index is zero whenever all the points are equivalent, no matter how many there are or how they are connected [5], e.g., $H(K_n) = H(C_n) = 0$ for all n.

(3.1.1) $$\chi_M(G) = [ne/(n + e)] \sum_{i<j} \sigma_{i,j}$$

(3.1.2) $$H = -\sum_i p_i \log_2 p_i$$

3.2 Complexity factors

Our approach abstracts the system to be studied as a graph G and then uses graph invariants $I(G)$ as indices of the complexity of G and the system it represents. At the very least, a complexity index must increase monotonically with those factors that contribute to complexity [2-4,37], viz. path length, branching, cycle size and *cyclization*, the number of cycles divided by the number of points. Some of these factors can be isolated in homologous series [2,4,36], e.g., path length in $\{P_n\}$, cycle size in $\{C_n\}$ and branching at a point in $\{K_{1,n}\}$.

Branching is dependent upon the degrees of the points, and it is often associated with the highest degree of the graph. Cyclization is impossible to isolate, since creating a new cycle by joining two points with a new line also increases the branching at both of the points. Branching and cyclization have long been recognized as complexity factors [9,37,38], and several branching indices [26-28,43,44] and cycle indices [45,46] have been devised. Multiple lines are a complicating factor discussed briefly in sections 2.2 and 3.6. Considering all the relevant factors, it can be concluded that $K_4 - x$ (bicyclobutane) is more complex than C_4 (cyclobutane); in section 2.3 these graphs were used to illustrate the calculation of a number of topological indices.

3.3 Homologous series

We discuss homologous series in some detail, as they are the keys to measuring the complexity of graphs and the molecules represented by them. Starting from a single point P_1, $\{P_n\}$ is the homologous series formed by joining a new vertex v to a point u of lowest degree in P_i with a new edge $x = uv$ to obtain P_{i+1}. Thus, P_{i+1} is the smallest supergraph of P_i that is also a path. Upon joining v to P_1, we obtain P_2, which has two

points of degree 1. Either may be considered the point of lowest degree for the next step, which gives P_3. In each subsequent step there are two points of degree 1 (the endpoints) and $n - 2$ points of degree 2. The paths $\{P_n\}$ represent the n-alkanes.

A second homologous series, the star graphs $\{K_{1,n}\}$, can be generated by starting with $K_{1,0} \cong P_1$ and adding a new vertex v to a point u of highest degree in $K_{1,i}$ with a new edge $x = uv$ to obtain $K_{1,i+1}$. In $K_{1,1} \cong P_2$ either point can be chosen, but thereafter we have one point of degree i and i points of degree 1. The homologous series $\{P_n\}$ and $\{K_{1,n-1}\}$ represent the least and most branched trees on n points, respectively.

A third homologous series is $\{C_n\}$, which can be constructed by connecting the endpoints of path P_n with an edge to give cycle C_n. The 1-cycle C_1 is defined as a loop connecting a vertex to itself (representing a lone-pair of electrons), the 2-cycle C_2 is the smallest multigraph with a double edge (representing a double bond), and the rest of the n-cycles are rings (representing cycloalkanes). This series is special, since the line graph is isomorphic to the graph, $L(C_n) \cong C_n$ ($n \geq 3$), which extends to the iterated line graphs as well.

The above homologous series add one new edge at a time. Starting with $K_1 \cong P_1$, the homologous series of complete graphs $\{K_n\}$ is generated by connecting a new vertex v to all i points of K_i with i new edges to obtain K_{i+1} at each stage. K_n is the most complex graph on n points from the viewpoint of branching and cyclization.

According to the definition of homologous series, one structural feature is repeatedly incremented; nevertheless, it is important to note that other structural features are also changed in the process. For example, in $\{P_n\}$ the number of vertices increases by 1 at each stage, and so does the number of edges, the number of connections η (paths of length 2) starting with P_3, the number of 'steric pairs' [27] (paths of length 3) starting with P_4, etc. Furthermore, $N_S = n$ and $N_T = n(n + 1)/2$. In $\{C_n\}$, $\eta = n$ for $n \geq 3$, $N_S = n + 1$ and $N_T = n^2 + 1$. In $\{K_{1,n}\}$, $\eta = n(n - 1)/2$, $N_S = n + 1$ and $N_T = 2^n + n$. In $\{K_n\}$, $\eta = n(n - 1)(n - 2)/2$. Rücker and Rücker have discovered formulas for N_S and N_T of K_n [47].

3.4 Hierarchies of homologous series

As discussed in the previous section, the cycle C_n can be derived from the path P_n by connecting the endpoints of the path with a new line. The definitions of C_1 and C_2 are novel; nevertheless, these graphs are demonstrably more complex than P_1 and P_2, respectively. Thus, C_n has the same number of points as P_n, but one more line and a new ring, which is a major complexity factor. Moreover, the degree of branching, another major complexity factor, increases at the points joined. (In C_1 the point is joined to itself.) We conclude that C_n is more complex than P_n for all n, and for the two homologous series we say that $\{C_n\}$ *dominates* $\{P_n\}$, symbolically $\{C_n\} > \{P_n\}$.

For an invariant $I(G)$ to be an intrinsic complexity index (cf. section 3.8), it must satisfy the inequality $I(C_n) \geq I(P_n)$. By an analogous argument, K_n is more complex than C_n for all $n > 3$, since it has $n(n-3)/2$ additional lines and more cycles. Thus, except for the first three graphs in each series, $\{K_n\} > \{C_n\}$ and $I(K_n) \geq I(C_n)$. Overall, we have the hierarchy $\{K_n\} > \{C_n\} > \{P_n\}$ $(n > 3)$ and the requirement that $I(K_n) \geq I(C_n) \geq I(P_n)$.

Another hierarchy for graphs on n points is $\{K_n\} > \{K_{1,n-1}\} > \{P_n\}$ $(n > 3)$, and $I(K_n) \geq I(K_{1,n-1}) \geq I(P_n)$ is also required for an intrinsic complexity index. It is easy to see why $\{K_n\} > \{K_{1,n-1}\}$ for $n \geq 3$; branching and cyclization are both greater in the former. (N.B., $K_1 \equiv K_{1,0}$ and $K_2 \equiv K_{1,1}$.) There is no universally accepted definition of branching, but there is agreement that branching depends on the degrees of the points. Since the sum of the degrees is constant for isomeric graphs, branching is primarily determined by the highest vertex degree. For $n > 3$ it is clear that $K_{1,n-1}$ and P_n are at the opposite extremes of complexity for trees on n points (cf. section 3.3), and $\{K_{1,n-1}\} > \{P_n\}$.

It should be noted that $I(K_{1,n-1}) \geq I(C_n)$ or $I(C_n) \geq I(K_{1,n-1})$ is not a requirement for a complexity index, as it would compare branching and cycle formation (i.e., 'apples and oranges'). Our previous indices η and N_T obey the above hierarchies; however, $C(\eta)$ does not, as $I(P_n) > I(C_n)$ for $n \geq 8$ (see Table 3).

3.5 New complexity indices

The simple invariant η is easy to count, but its discriminating power is limited. This situation can be improved somewhat by calculating $C(\eta)$; however, it does not obey all the hierarchies (previous section). Index N_T is highly discriminating, but difficult to determine for large structures. In order to develop robust new complexity indices, we investigated selected subgraphs, such as K_i (cliques) and $K_{i,j}$ (bicliques). Several promising new complexity indices are discussed below. For indices that describe the complexity of homologous series, continuously increasing (discrete) functions are preferred, although monotonic non-decreasing functions are allowed.

The indices collected in Table 1 are the intersection number $i(G)$, the bintersection number $j(G)$, and the corresponding functions for the line graphs $L(G)$ and $L^2(G)$. Table 2 summarizes the number of kinds of minimal edge clique covers k_S and the total number of minimal edge clique covers k_T. It also contains the corresponding indices k_S^{bi} and k_T^{bi} for minimal edge biclique covers. The numbers of kinds of partitions p_S and p_S^{bi} and their total numbers p_T and p_T^{bi} for cliques and bicliques, respectively, are also of interest, and the values for bicliques are compiled in Table 2. These invariants are defined rigorously in section 2.2, and sample calculations are given in section 2.3.

Index $i(G)$ is a continuously increasing discrete function for the homologous series $\{P_n\}$, $\{C_n\}$ and $\{K_{1,n}\}$ with a few well-defined exceptions for small graphs. Thus, $P_1 \cong K_{1,0} \cong K_1$, $P_2 \cong K_{1,1} \cong K_2$ and $C_3 \cong K_3$ are complete graphs, and $i(K_n) = 1$ for all n. Thus, $i(G)$ is the constant function 1 for $\{K_n\}$. Finally, this invariant gives the wrong order of complexity for C_4 and $K_4 - x$ (section 2.3). It may have limited usefulness in some applications; however, the intersection number $i(G)$ is not a robust complexity index.

With the exception of K_1, K_2 and K_3, the line graphs of K_n are not complete graphs; therefore, $i[L(K_n)]$ is not the constant function 1. However, the line graphs of the star graphs are complete graphs; consequently, $i[L(K_{1,n})] = i(K_n) = 1$. Thus, the intersection number of the line graph, $i[L(G)]$, is also not a robust complexity index.

If we take the intersection number of the second line graph, $i[L^2(G)]$, then all the sequences for the homologous series in Table 1 are monotonic non-decreasing functions. In fact, if the smallest two or three values for each sequence are ignored, $i[L^2(G)]$ is a continuously increasing discrete function for all of the homologous series. Finally, $i[L^2(G)]$ obeys all the hierarchies discussed above (section 3.4), and we conclude that it is a good index of intrinsic complexity.

With one exception, the bintersection number $j(G)$ is a monotonic non-decreasing function for each of the homologous series in Table 1, as are $j[L(G)]$ and $j[L^2(G)]$. The exception is $\{C_n\}$, which has a dip at $C_4 \cong K_{2,2}$ for these indices, since $j(K_{m,n}) = 1$ for all m and n. Furthermore, $j(G)$ is the constant function 1 for the star graphs $\{K_{1,n}\}$. For some of the homologous series, indices $j[L^n(G)]$ are step functions, which are generally not as useful as continuously increasing ones. They may be useful as complexity measures in some applications, but are not robust indices.

Turning to Table 2, the invariants k_S and k_T are clearly not useful, even as rough estimates of complexity, owing to complete degeneracy in $\{P_n\}$, $\{C_n\}$ (except for C_3), and $\{K_{1,n}\}$. With the exception of $p_S(K_4) = 3$ and $p_T(K_4) = 6$, p_S and p_T are the same as the respective k-indices for the graphs in Table 2, and consequently they are not listed.

On the other hand, the bipartite analogs k_S^{bi} and k_T^{bi} appear to be good complexity indices, and so does p_T^{bi}. First and foremost, these indices obey all the hierarchies. Except for the first entries in $\{P_n\}$, $\{K_{1,n}\}$, and $\{K_n\}$ and one degeneracy in $\{C_n\}$, these indices are continuously increasing discrete functions. Owing to the presence of an extra minimal edge biclique cover for $C_4 \cong K_{2,2}$, the value of k_S^{bi} for C_4 is the same as for C_5. Graphs $P_1 \cong K_{1,0} \cong K_1$ and $P_2 \cong K_{1,1} \cong K_2$ are an isolated point and a single line with its endpoints, respectively, and may be considered the primitives of a graph. It is not a serious flaw for an invariant to have the same value for these trivial graphs. The index p_S^{bi} does not obey all the hierarchies, e.g., its value is 12 for P_8, but only 8 for C_8.

For the homologous series $\{P_n\}$ and $\{K_{1,n}\}$, in which one new vertex and one new edge are added at each stage, p_T^{bi} can be calculated from familiar counting sequences. The

values of $p_T^{bi}(P_n)$ are given by the Fibonacci numbers, and the values of $p_T^{bi}(K_{1,n})$ are given by the Stirling numbers of the second kind [19]. These sequences are the limits for the least branched and most branched trees, respectively (cf. section 3.3).

Index p_T^{bi} for $\{C_n\}$ is more complicated to calculate. Starting from C_3 and except for C_4, it is the sum of three identical Fibonacci sequences, one of them offset. The graph C_4 is the only one of the n-cycles ($n \geq 3$) that has $K_{2,2}$ as a partition; all the rest have partitions that contain exclusively $K_{1,2}$ and $K_{1,1}$. Subtracting the unique partition from p_T^{bi}, we have $p_T^{bi}(\text{corr}) = 7$ for C_4. With this correction, $p_T^{bi}(C_n, \text{corr})$ is given by equation 3.5.1 for $n \geq 3$, where the Fibonacci sequences start with $p_T^{bi}(P_0) = 0$ and $p_T^{bi}(P_1) = 1$.

(3.5.1) $$p_T^{bi}(C_n, \text{corr}) = 2\, p_T^{bi}(P_n) + p_T^{bi}(P_{n-3})$$

3.6 Extensions

The extension of the new complexity indices to multigraphs is straightforward. For example, the unbranched 1-alkenes can be derived from the n-alkanes (P_n) by adding another edge joining vertices v_1 and v_2. Arguments analogous to the ones in section 3.4 give {1-alkenes} > {n-alkanes} for these homologous series, and the new intrinsic complexity indices obey I(1-alkene) $\geq I$(n-alkane). The simplest multigraph is C_2, the graph of ethylene, which has $k_T = 1$, $k_T^{bi} = 3$, $i(C_2) = 2$ and $j(C_2) = 1$, following our definitions. For unbranched alkenes $K_{1,1}$ and $K_{1,2}$ are the only bicliques; $K_{1,3}$ and $K_{1,4}$ are possible for branched alkenes, as for branched alkanes.

Several variations on k_S^{bi} and k_T^{bi} are possible. The delimiter "minimal" can be dropped and all possible edge covers or set system covers can be considered (see section 2.3 for examples). Note that the total number of minimal edge biclique covers grows very fast, so that they are not yet counted for K_4. The total numbers of covers and partitions considered thus far assume labeled graphs (e.g., $k_T^{bi} = k_L^{bi}$, $p_T^{bi} = p_L^{bi}$). The number k_U^{bi} of minimal edge biclique covers for an unlabeled graph can also be used, as illustrated for $K_4 - x$ in Table 4, where $k_U^{bi} = 33$. For a given graph $k_S^{bi} \leq k_U^{bi} \leq k_T^{bi}$.

The 'all possible subgraphs' method (N_T and N_S with connected subgraphs) was originally introduced to measure the similarity of molecules, although its relevance to complexity was noted in the first paper on this approach [48]. The edge covers and partitions provide an alternative method to measure similarity.

3.7 Composite indices

One approach to compensating for the deficiencies of two or more topological indices is to take a weighted average [44]. For example, the sum $s = i(G) + j(G)$ is a monotonic non-decreasing function for each of the homologous series. The problem with such composite indices is the way they combine the different complexity factors can lead to inconsistencies. The value of this index for the path P_5 is $s = 6$, which is higher than its value $s = 5$ for the star graph $K_{1,4}$ with the same number of points. However, the latter is more branched and consequently more complex, since there are no other complicating factors. It must be concluded that $i(G) + j(G)$ is not an intrinsic complexity index, since it violates the requirement that $I(K_{1,n-1}) \geq I(P_n)$.

Another approach to the evaluation of complexity involves the partial order that results when two or more indices are used simultaneously [49,50], but maintain their individual identities. The partial order is induced on a grid embedded in the appropriate multi-dimensional real space, which is viewed as a Hasse diagram. The multi-dimensional grid has graphs as its vertices, and the coordinates of a graph are determined by the values of its indices, which are each plotted on a different axis.

In our case the conventional direction of a Hasse diagram is SW (map directions), as the lines joining vertices are arcs from N to S or E to W. Then, the vertex in the lower left corner is the smallest, while vertices become increasingly larger in the direction E (to the right), N (upwards) or NE. Adjacent vertices are always comparable ($N \geq S$ and $E \geq W$), and a vertex is considered comparable to itself. Non-adjacent vertices are *not comparable* when a new line connecting them would have a NW or SE direction; e.g., the vertex at coordinates (2,1) is not comparable to the vertex at (1,2).

In accord with our conventional direction, the vertex at (2,2) is greater than the vertex at (1,1). We can draw an analogy to vector addition, where the sum of two vectors placed head to tail is the vector that extends from the tail of the first to the head of the second. Since the lines joining vertices in the Hasse diagram are arcs from N to S or E to W, the only valid resultant vectors point SW. (N.B., the S and W components need not be equal in length.) For posets to properly represent complexity, the hierarchies of homologous series must be obeyed, and the arcs or vectors from the more complex graph to the less complex one must point S, W or SW.

For selected pairs of the indices, we construct 2-dimensional grids and the posets they define. The graphs used are the ones in the homologous series $\{P_n\}$, $\{C_n\}$, $\{K_{1,n}\}$ and $\{K_n\}$, and the indices tested are from Table 1. We first consider the indices $i(G)$ and $j(G)$, as illustrated in Figure 1. In this poset P_2 is W of C_2, which is the correct direction; however, P_3 is SE of C_3 and P_4 is NW of C_4, so that the graphs in these pairs are non-comparable. Thereafter, P_n is SW of C_n (odd n) or P_n is W of C_n (even n), both of which are consistent with $I(P_n) \leq I(C_n)$. The direction is wrong for the pairs of graphs P_4 and $K_{1,3}$, P_5 and $K_{1,4}$, and P_6 and $K_{1,5}$. Thus, $i(G)/j(G)$ does not order graphs according to intrinsic complexity (see also next section).

Next the partial order induced by $i[L(G)]$ and $j[L(G)]$ is examined (Figure 2). In this poset P_n is S, W or SW of C_n for all $n \geq 3$. However, P_5 is E of $K_{1,4}$, which is the wrong direction. The pairs of graphs P_4 and $K_{1,3}$ and P_6 and $K_{1,5}$ are non-comparable. All things considered, the Hasse diagram for $i[L(G)]/j[L(G)]$ is also not suitable for treating intrinsic complexity.

The Hasse diagram for $i[L^2(G)]/j[L^2(G)]$ is presented in Figure 3. In this poset P_3 is S of C_3. Furthermore, P_4 is W of C_4, and both of these graphs are SW of K_4. Also, P_4 is S of $K_{1,3}$, and they are both SW of K_4. Finally, P_5 is SW of both C_5 and $K_{1,4}$, and P_6 is SW of both C_6 and $K_{1,5}$. All these directions are correct. Within each homologous series all pairs of graphs have the correct orders, except for C_3 and C_4, which are non-comparable. It appears that all the hierarchies are obeyed, and the Hasse diagram for $i[L^2(G)]/j[L^2(G)]$ is a valid way to look at intrinsic complexity.

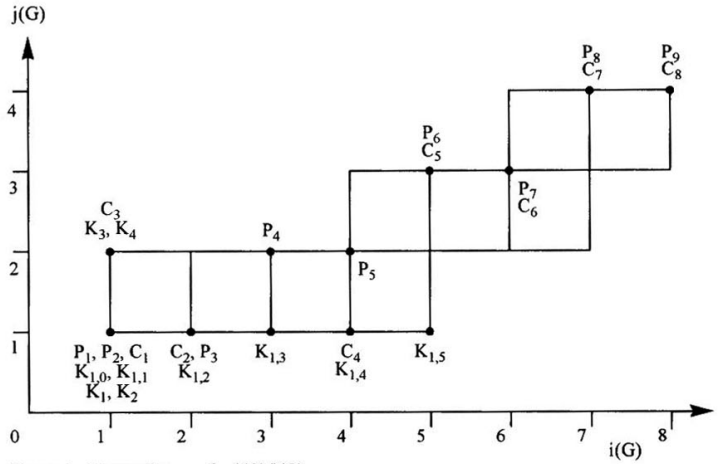

Figure 1. Hasse diagram for i(G)/j(G).

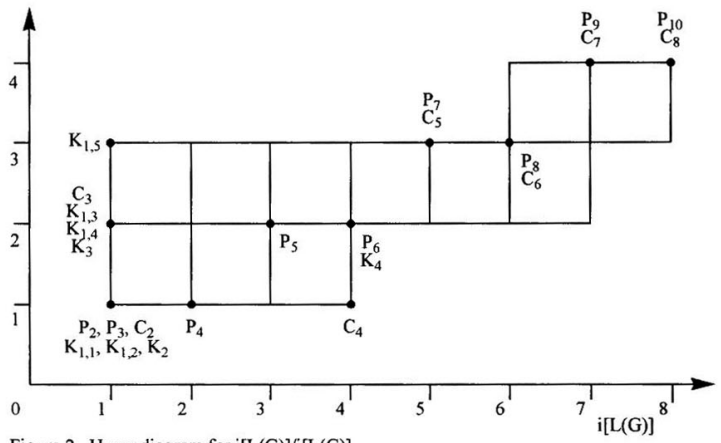

Figure 2. Hasse diagram for i[L(G)]/j[L(G)].

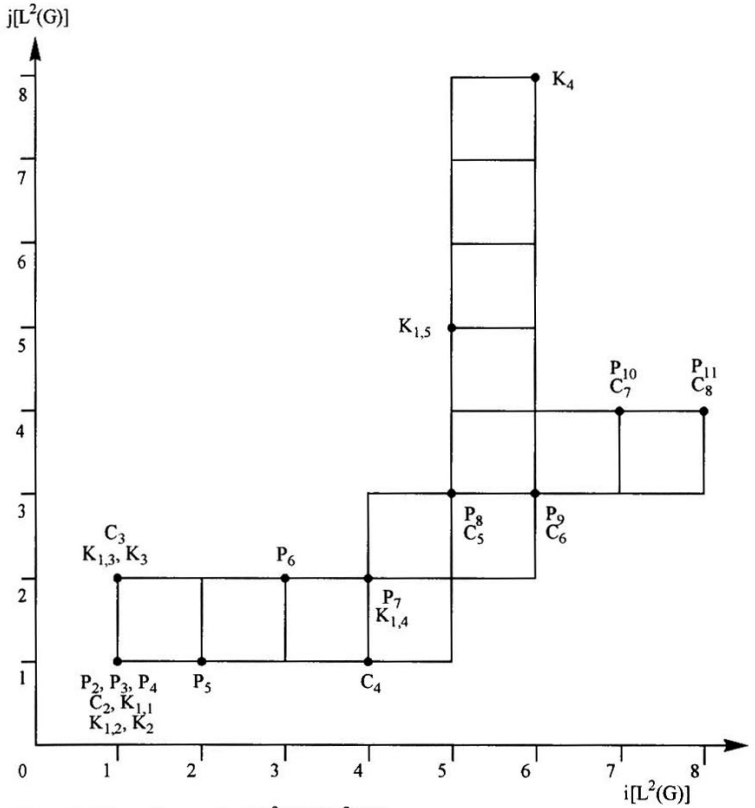

Figure 3. Hasse diagram for $i[L^2(G)]/j[L^2(G)]$.

3.8 Complexity and symmetry

As mentioned in the introduction, the two main approaches to understanding the complexity of an object (the 'system') are (i) to examine the way in which it was constructed, i.e., assembled from its parts, which can be as small as atoms and subatomic particles or as large as stars and galaxies, and (ii) to examine the way in which it operates, i.e., the way its parts interact with each other.

The problem with the former approach is we often do not know how an object was constructed. The best we can do is some sort of reverse engineering to get an idea of how it might have been assembled. Moreover, several ways of making the same object may be possible, and they depend upon the tools available. Thus, the complexity of construction is not an intrinsic property. A variation on the construction theme is to calculate the minimum information required for instructions on how to make the object, which can be thought of as constructing it conceptually. An example of conceptual construction is the 'minimum part and placement' description, which has even been applied to the biological cell [8]. The complexity of conceptual construction is not intrinsic, either, as it implicitly assumes a tool set and different instruction sets are possible. Thus, we can consider the complexity of constructing the physical model or the corresponding conceptual model [51]. We classify both of these variations under *extrinsic complexity*. To differentiate them, we call the complexity of actually fabricating an object the *constructive extrinsic complexity* or simply the *constructive complexity*.

Intrinsic complexity describes the interaction approach (ii), and we measure it by using graph invariants. We call it intrinsic because the interactions of its parts are unique for any individual object. The question then arises whether we can distinguish the interactions of the physical model from those of the conceptual model, and the short answer is 'no.' A conceptual model is created when we abstract the system as a graph in the first step of our method. We could attempt to measure the complexity of the interactions in the system by using an information approach, but then it would be similar to extrinsic complexity, as it would depend on the code or language used.

The pervasive opinion persists that symmetry decreases complexity because it reduces the necessary information [11,12,52], which is an extrinsic view. Ergo, a highly symmetrical graph such as K_n is simple, since it is possible to describe one of its points as having degree $n - 1$ and then duplicate it n times. Note that C_n is equally complex—or simple—in this scheme, which is clearly related to the complexity of conceptual construction. When defined according to the Rashevsky-Mowshowitz method (equation 3.1.2), the complexity of K_n, C_n, and all other regular graphs is the same—zero.

A way to appreciate the complexity of K_n versus C_n is to consider the problem of drawing them on paper, which is equivalent to actually constructing the physical model. In both cases the pen, whether in a plotter or someone's hand, must draw n points; however, in the former it draws $n(n - 1)/2$ lines, whereas in the latter it only draws n lines. Assuming a constant rate of work and a constant flow of ink, the time, materials and energy required for K_n are ca. $(n - 1)/2$ times those for C_n. While symmetry may simplify the conceptual construction, it does not simplify the actual fabrication or decrease the constructive complexity. We can also consider the pen to trace out the interactions between the points (approach ii), which illustrates the close relationship that can exist between constructive complexity and intrinsic complexity (cf. section 3.9).

In another example, graphs can represent communications networks, where the points stand for people or telephones or computers and the lines for the links between them. Following the usual logic of extrinsic complexity, the complete graph has been identified as the simplest connected communications network [11]. However, if one is confronted by the costs and logistics of actually building the network or the task of routing messages through it and measuring its performance, the problem is not so simple [53].

3.9 Molecular complexity and molecular symmetry

As demonstrated in the previous section, a distinction can be drawn between the constructive complexity of actually making a physical object and the complexity of describing it at a level of detail that would enable one to make it, taking the description as instructions. When the object is a molecule, we have discussed *synthetic complexity*

[2,3], which is synonymous with constructive complexity. We have also defined molecular complexity according to the information approach [4,5,38], i.e., *extrinsic molecular complexity*, and molecular complexity according to the interaction approach [2,36,37], i.e., *intrinsic molecular complexity*. An example of the former is the length of a standard linear notation [54]; η and N_T are measures of the latter.

The synthetic complexity of molecules has not been quantified, as it depends on the tools available, i.e., the state of the art of synthetic chemistry. The tools of synthetic chemistry are the starting materials and the reactions that convert them into products. Reducing these considerations to a single number is not practical at this time. We have focused our studies on molecular complexity [2-4] and the synthetic reactions that produce the greatest increases in it [3,36]. The difference between synthetic complexity and intrinsic molecular complexity can be appreciated by noting that, as the state of the art evolves with time, the synthetic complexity of making a given molecule may change, but the intrinsic molecular complexity of the molecule itself does not, since the structure remains the same.

The extrinsic complexity of conceptual construction can be thought of as the ideal limit of constructive complexity, which is the extrinsic complexity of actual construction (cf. section 3.8). They are equal at a state of the art where all possible tools are available, including the tools needed to make use of any symmetry present. Synthetic chemistry is approaching this state asymptotically as new synthetic reactions are discovered, invented or developed. An index of complexity is useful as a 'yardstick' to judge the state of the art; e.g., we have calculated the changes in molecular complexity that accompany old and new synthetic reactions [3,38].

It has often been assumed that molecular symmetry is a simplifying factor in chemical synthesis [55]; however, it can be a complicating factor if a reaction that makes it in a target or breaks it in a starting material is needed and cannot be found [24,56]. Thus, equation 2.2.2 for $C(\eta)$, which is sensitive to symmetry, is proportional to synthetic complexity in an 'ideal world' with all possible synthetic reactions, including symmetry-breaking 'Odin' reactions [57].

Symmetry is effective when it simplifies the synthesis graph, e.g., when there is molecular symmetry and also a reaction that makes efficient use of it. It is important to note that molecular symmetry is not necessary to have synthesis graph symmetry, which we termed *reflexivity* [24]. A substructure (abstracted as a subgraph) may be repeated two or more times in such a way that the copies (isomorphic subgraphs) are not related by any element of molecular symmetry. If reactions can be found that incorporate a precursor molecule multiple times, great economies can be realized, as reflected in a simpler synthesis graph [24]. In the absence of reflexivity, the effects of molecular complexity on the evaluation of a synthesis plan are more subtle, and we introduced the 'complexity vs. step' plot to visualize how the molecular complexity changes during the course of a synthesis [58,59,60].

4. Conclusion

The kinds of complexity have been clarified, and they are classified as intrinsic and extrinsic. Intrinsic complexity is associated with the interactions in a system. By mathematically formalizing the concept of the homologous series, borrowed from organic chemistry, we are able to test graph invariants $I(G)$ against archetypal homologous series $\{P_n\}$, $\{C_n\}$, $\{K_{1,n}\}$ and $\{K_n\}$ to determine which invariants are intrinsic complexity indices. It is necessary, but not sufficient, for indices of intrinsic complexity to increase monotonically within these homologous series. They must also be consistent with hierarchies of homologous series, embodied in the inequalities $I(K_n) \geq I(K_{1,n-1}) \geq I(P_n)$ and $I(K_n) \geq I(C_n) \geq I(P_n)$. Branching is reflected in $\{K_{1,n-1}\} > \{P_n\}$ and the effect of cyclization is manifested, albeit admixed with branching, in $\{K_n\} > \{K_{1,n-1}\}$, $\{K_n\} > \{C_n\}$, and $\{C_n\} > \{P_n\}$. We have investigated several new graph invariants and discovered that some of them are good intrinsic complexity indices, viz. the intersection number of the second line graph, $I[L^2(G)]$, and certain invariants based on biclique covers: k_S^{bi}, k_T^{bi} and p_T^{bi}. Individual indices can be combined into composite indices in two ways. Either the individual indices lose their identities in a sum or they keep them in a poset, which can be studied with the aid of a Hasse diagram. The role of symmetry has also been investigated; it appears to be a simplifying factor for extrinsic complexity, but not for intrinsic complexity.

References

[1] This is paper 15 in a series on molecular complexity and applications of graph theory to chemistry. For paper 14 see ref. 2.

[2] S.H. Bertz and W.F. Wright, Graph Theory Notes of New York (NY Acad. Sci.) **XXXV** (1998) 32-48.

[3] S.H. Bertz and T.J. Sommer, pages 67-92 in *Organic Synthesis: Theory and Applications*, vol. **2**, Ed. T. Hudlicky (JAI Press, Greenwich, CT, 1993).

[4] S.H. Bertz, pages 206-221 in *Chemical Applications of Topology and Graph Theory*, Ed. R.B. King (Elsevier, New York, 1983).

[5] S.H. Bertz, Bull. Math. Biol. **45** (1983) 849-855.

[6] S. Lloyd and H. Pagels, Ann. Phys. **188** (1988) 186-213.

[7] F. Papentin, J. Theor. Biol. **87** (1980) 421-456; **95** (1982) 225-245.

[8] R. Hinegardner and J. Engelberg, J. Theor. Biol. **104** (1983) 7-20.

[9] D. Bonchev and N. Trinajstić, J. Chem. Phys. **67** (1977) 4517-4533.

[10] L.B. Kier and B. Testa, Adv. Drug Res. **26** (1995) 1-43.

[11] M. Gell-Mann, *The Quark and the Jaguar: Adventures in the Simple and the Complex* (W.H. Freeman, New York, 1994).

[12] J.L. Casti, *Complexification* (Harper Collins, New York, 1994).

[13] R. Badii and A. Politi, *Complexity: Hierarchical Structures and Scaling in Physics* (Cambridge University Press, Cambridge, 1997).

[14] a) D.J. Klein and D. Babić, J. Chem. Inf. Comput. Sci. **37** (1997) 656-671; b) D.J. Klein, J. Math. Chem. **18** (1995) 321-348.

[15] G. Birkhoff, *Lattice Theory* (American Mathematical Society, Providence, RI, 1948).

[16] K.A. Ross and C.R.B. Wright, *Discrete Mathematics*, 4th ed. (Prentice-Hall, Upper Saddle River, NJ, 1999).

[17] F. Harary, *Graph Theory* (Addison-Wesley, Reading, MA, 1969).

[18] T.A. McKee and F.R. McMorris, *Intersection Graph Theory* (SIAM, Philadelphia, PA, 1999).

[19] D.I.A. Cohen, *Basic Techniques of Combinatorial Theory* (Wiley, New York, 1978).

[20] L.W. Beineke, pages 17-23 in *Beiträge zur Graphentheorie*, Ed. H. Sachs, H. Walther and H.-J. Vose (Teubner-Verlag, Leipzig, 1968).

[21] V.V. Menon, Amer. Math. Monthly **73** (1966) 986-989.

[22] R.A. Bari, Ann. Discrete Math. **13** (1982) 15-22.

[23] S.H. Bertz, J. Chem. Soc., Chem. Commun. (1986) 1627-1628.

[24] S.H. Bertz, J. Chem. Soc., Chem. Commun. (1984) 218-219.

[25] N. Trinajstić, *Chemical Graph Theory*, 2nd ed. (CRC Press, Boca Raton, FL, 1992).

[26] M. Gordon and J.W. Kennedy, J. Chem. Soc., Faraday Trans. II **69** (1973) 484-504, and references cited therein.

[27] J.R. Platt, J. Phys. Chem. **56** (1952) 328-336, and references cited therein.

[28] S.H. Bertz, Discrete Appl. Math. **19** (1988) 65-83.

[29] *Topological Indices and Related Descriptors in QSAR and QSPR*, Ed. J. Devillers and A.T. Balaban (Gordon & Breach, New York, 1999).

[30] D. Bonchev, SAR & QSAR Environ. Res. **7** (1997) 23-43.

[31] S.C. Basak and V.R. Magnuson, Discrete Appl. Math. **19** (1988) 17-44.

[32] P. Erdös, A. Goodman and L. Posa, Can. J. Math. **18** (1966) 106-112.

[33] L.W. Beineke and C.M. Zamfirescu, Discrete Math. **39** (1982) 237-254.

[34] E.S. Klein and C.M.D. Zamfirescu, Congressus Numerantium **110** (1995) 137-144.

[35] M. Sen, S. Das, A.B. Roy and D.B. West, J. Graph Theory **13** (1989) 189-202.

[36] S. H. Bertz and T.J. Sommer, Chem. Commun. (1997) 2409-2410.

[37] S.H. Bertz, J. Chem. Soc., Chem. Commun. (1981) 818-820.

[38] S.H. Bertz, J. Am. Chem. Soc. **103** (1981) 3599-3601.

[39] N. Rashevsky, Bull. Math. Biophys. **17** (1955) 229-235.

[40] A. Mowshowitz, Bull. Math. Biophys. **30** (1968) 175-204.

[41] a) R.L. Brooks, C.A.B. Smith, A.H. Stone and W.T. Tutte, Duke Math. J. **7** (1940) 312-340. See also b) I. Gutman, R.B. Mallion and J.W. Essam, Mol. Phys. **50** (1983) 859-877.

[42] D. Minoli, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (Ser. 8) **59** (1975) 651-661.

[43] M. Randić, J. Am. Chem. Soc. **97** (1975) 6609-6615.

[44] H.P. Schultz, J. Chem. Inf. Comput. Sci. **29** (1989) 227-228.

[45] M. Randić, J. Chem. Inf. Comput. Sci. **37** (1997) 1063-1071.

[46] D. Bonchev, A.T. Balaban, X. Liu and D.J. Klein, Int. J. Quantum Chem. **50** (1994) 1-20.

[47] G. Rücker and C. Rücker, Unpublished Results (C. Rücker, letter of 17 January 2000 to S.H. Bertz).

[48] S.H. Bertz and W.C. Herndon, pages 169-175 in *Artificial Intelligence Applications in Chemistry* (ACS Symposium Series vol. **306**), Ed. T.H. Pierce and B.A. Hohne (American Chemical Society, Washington, DC, 1986).

[49] M. Randić, J. Chem. Edu. **69** (1992) 713-718, and references cited therein.

[50] D. Bonchev, O. Mekenyan and N. Trinajstić, J. Comput. Chem. **2** (1981) 127-148.

[51] C.J. Suckling, K.E. Suckling and C.W. Suckling, *Chemistry through Models* (Cambridge University Press, Cambridge, 1978).

[52] J. Rosen, *Symmetry Discovered: Concepts and Applications in Nature and Science* (Cambridge University Press, Cambridge, 1975).

[53] J.L. Hennessy and D.A. Patterson, *Computer Organization and Design*, 2nd ed. (Morgan Kaufmann, San Francisco, 1998).

[54] W.C. Herndon and S.H. Bertz, J. Comput. Chem. **8** (1987) 367-374.

[55] T.-L. Ho, *Symmetry: A Basis for Synthesis Design* (Wiley, New York, 1995).

[56] S.H. Bertz, Tetrahedron Lett. **24** (1983) 5577-5580.

[57] D.J. Milner, J. Chem. Tech. Biotechnol. **63** (1995) 301-312.

[58] S.H. Bertz, J. Am. Chem. Soc. **104** (1982) 5801-5803.

[59] F. Serratosa, *Organic Chemistry in Action: The Design of Organic Synthesis* (Elsevier, Amsterdam, 1990) 19-22, 67, 331.

[60] M. Chanon, R. Barone, C. Baralotto, M. Juillard and J.B. Hendrickson, Synthesis (1998) 1559-1583.

Acknowledgments

We thank A.T. Balaban, D.J. Klein and C. Rücker for many helpful suggestions. W.F. Wright provided references and S. D'Arcangelis assisted with the figures. This work was supported in part by the PCS-CUNY Award Program.