Quantitative Measures of Network Complexity

CHAPTER · MAY 2007

DOI: 10.1007/0-387-25871-X_5

2 AUTHORS, INCLUDING:

Danail Bonchev
Virginia Commonwealth University
273 PUBLICATIONS **4,281** CITATIONS

Chapter 5

Quantitative Measures of Network Complexity

Danail Bonchev and Gregory A. Buck

Center for the Study of Biological Complexity

Virginia Commonwealth University

Richmond, Virginia 23284-2030

dgbonchev@vcu.edu

5.1. Some History

The first attempts to evaluate quantitatively the complexity of a system have been related to complexity of cells, organisms, and humans. Fascinated by the complex nature of the living things, a group of young mathematical biologists applied in the 1950s the Shannon theory of communications[1] to assess the information content of the living matter.[2-5] The analysis made by Rashewsky[4] provided the first proof that life on earth cannot emerge as a random event, because the probability for such an event would be incredibly small. Two different approaches have been used in defining the information content. The first one proceeded from the elemental composition of the living matter (C, N, O, etc.) and is the predecessor of what is nowadays called *compositional complexity*. Rashewsky' *topological information* has been based on partitioning the atoms in a structure according to both their chemical nature and their equivalent topological neighborhoods. Mowshovitz[6] developed further these ideas to define complexity of graphs. Minoli[7]

introduced his *combinatorial complexity* of graphs, proceeding from the count of the graph vertices, edges, and paths.

In parallel with these attempts, another definition of information content has been advanced by Kolmogorov.[8] His *algorithmic information* has been defined as the minimal length of the program that exhaustively describes a given system. This type of information measure has found a broad application in computer sciences. The relevance of algorithmic information in describing *structural* complexity, however, is low,[9] which limited its application to chemistry, whereas in biology it has found some application in assessing the genome complexity.

Shannon's information has been widely applied in chemistry in the form of information indices, characterizing different aspects of chemical structure.[10-16] These structural descriptors have been commonly used for quantitative structure-property and structure-activity relationships (QSPR and QSAR). However, only few of them have satisfied the requirements for a complexity measure.[17] Bertz introduced in 1981 his molecular complexity index applying Shannon's equation to the distribution of the two-edge subgraphs in molecular graphs.[18] That was the starting point of a systematic search in chemical theory for relevant measures of molecular complexity, a search that shifted the focus from information theory to molecular topology and graph theory. A series of requirements have been formulated for a structural descriptor to be a complexity measure,[19-21] along with hierarchical concepts of molecular complexity.[22,23] A number of high quality measures of topological complexity have been devised during the last 7-8 years.[24-31] Complexity of chemical reaction networks has also been addressed making use of the spanning subgraphs of these cyclic graphs. [32-35]

In the meantime, in the middle of 1980s, complexity theory emerged as a new integrative branch of science. The emphasis in the new theory was put on the complex dynamic systems, systems characterized by nonlinear dynamics and emergent events. The quantitative aspects of the theory, related to random graphs, did not bring exciting results. The situation changed radically only when it was realized that any dynamic evolutionary system could be adequately presented by a network (a graph) that is non-random. Thus, complexity theory has found its universal language to describe systems as diverse as discrete space-time, the living cell, ecosystems, financial markets, World Wide Web, and social systems. This opened the door to the introduction of general methods for characterizing systems complexity, not only as information-based compositional complexity but, most essentially, as topological complexity of the network representing the system.

This chapter aims at elucidating the methods for quantitative assessments of networks complexity. It borrows from the rich arsenal of such methods developed during the last 25 years in chemical graph theory and chemical information theory. Being devised in a sophisticated way so as to distinguish the complexity of the multitude of molecules, these methods will be presented in a form adapted to the very large size of networks in biology and ecology. New graph invariants having properties of complexity measures will also be presented. Examples of cellular and ecological networks will be analyzed with the methods presented.

5.2. Networks as Graphs

Networks are well characterized both quantitatively and as structural patterns or motifs by graph theory, which has at least 150 years of extensive development and application. Graph theory as a branch of discrete mathematics has been brought to life to solve specific problems from three different areas of science. Leonard Euler in 1788 constructed the first graph to solve the famous mathematical puzzle for the Königsberg bridges, a problem that is a predecessor of the transport and communication sets problems of our time. Rudolf Kircchoff in mid 19th century reinvented graphs and developed their theory to solve fundamental problems of electrical sets, a work of great value for the electronic networks of the 21st century, as well as for the complex chemical reaction networks. The third root of graph theory is in structural chemistry, which in the last part of 19th century was trying to determine the number of isomers, chemical compounds having the same atomic composition but different spatial structure.
The variety in the graph theoretical background produced a variety of non-standardized terminologies. In this chapter, we shall follow mainly the manner the terminology is used in chemical graph theory. Cellular networks are molecular networks, and we believe that the use of terms like "wirings" coming from electrical and computer engineering should be avoided in describing living things. This section introduces some basic graph theoretical notions and descriptors needed for the network topological and complexity analysis.

5.2.1. Basic Notions in Graph Theory[36-38]

A network is defined by the set of V *vertices* (nodes, points), $\{V\}\equiv\{v_1, v_2, \dots, v_V\}$, and the set of E *edges* (links, lines), $\{E\}\}\equiv\{E_1, E_2, \dots, E_E\}$. The edge $\{ij\}$ is the line that emanates from vertex i and ends in vertex j. A *subgraph* is a graph obtained from the parent graph by deleting at least one edge or a vertex with its incident edges. A *loop* is an edge that begins and ends in the same vertex. A *multigraph* is a graph in which some pairs of vertices are linked by more than one edge. *Simple graphs* are graphs having no multiple edges and loops. In a *complete graph*, K_V, any two vertices are connected by an edge. A *directed* graph is a graph having at least one directed edge. Directed edges are termed *arcs*. Graph without any directed edge is *undirected.* The graph is *connected* when there is a path between any pair of vertices in it; otherwise the graph is *disconnected*. A *path* in the graph is a sequence of adjacent edges without traversing any vertex twice. A *path graph*, P_V, is a graph containing only one path. A *star*-graph, S_V, is a graph containing one central vertex and V-1 *branches* of length one edge. A *walk* is an alternating sequence of vertices and edges, each of which could be traversed more than once. The *walk length* is the number of edges in it. A *cycle* is a path that starts from and ends in the same vertex. Graphs containing at least one cycle are called *cyclic graphs. Trees* are graphs containing no cycles. A *spanning tree* is a connected acyclic graph containing all the vertices of the graph. Graph *components* are connected subgraphs or vertices that are not connected to each other. Euler's theorem relates the number of vertices V, edges E, independent cycles C, and components K:

$$C = E - V + K \tag{1}$$

Fig.1 illustrates the notions introduced.

5.2.2. Adjacency Matrix and Related Graph Descriptors

Two vertices *j* and *i* are called *adjacent* when they are connected by an edge {i,j}. The adjacency relation is quantified by the term $a_{ij} = 1$, and the no adjacency one by $a_{ij} = 0$. The number of the nearest-neighbors of a vertex *i* is termed *vertex degree*, a_i. *Vertex degree distribution* is an ordered, usually descending set of vertex degrees, $\{V_{ord}\}\equiv\{v_{max}, \ldots, v_{min}\}$.The sum of all vertex degrees in a graph defines its *total adjacency*, ***A.*** The matrix containing all adjacency relations in a graph *G* is called *adjacency matrix*, $\mathbf{A}(G)$. The vertex degree of vertex *i* is calculated as the sum over all entries in the i^{th} row of adjacency matrix. Similarly, the total adjacency of graph *G*, $A(G)$, is calculated also as the sum over all matrix elements, a_{ij}:

$$a_i = \sum_{j=1}^{V} a_{ij}\ ; \quad A(G) = \sum_{i=1}^{V}\sum_{j=1}^{V} a_{ij} = \sum_{i=1}^{V} a_i \tag{2a,b}$$

Undirected graphs (*G*) have adjacency matrices that are symmetrical with respect to their main diagonal, $a_{ij} = a_{ji}$. In directed graphs (*DG*), the symmetry of adjacency matrix is destroyed. Examples are shown in Fig. 2.

The vertex degrees of graph **2** shown above are actually *out-degrees*; they count the outgoing edges but not the incoming ones. Similarly, $A(\mathbf{2})$ shown is *out-adjacency*, $A_{out}(\mathbf{2})$, and the vertex degree distribution is an *out*-degree one {3, 3, 2, 2, 1, 0}. Vertices 4, 5, and 6 have also *in-degrees* equal to 1, thus defining a vertex *in*-degree distribution {1, 1, 1, 0, 0, 0}, and producing $A_{in}(\mathbf{2}) = 3$. One may generalize that the sum of the *in*- and *out*-adjacencies of the directed graph are equal to the adjacency of the parent undirected graph:

$$A_{out}(DG) + A_{in}(DG) = A(G) \tag{3}$$

The adjacency matrix of a graph provides also some generalized descriptors of network connectivity like the *average vertex degree* $<a_i>$ and *connectedness* (or connectance), *Conn*:

$$< a_i > = \frac{A}{V}; \qquad Conn = \frac{A}{V^2} = \frac{2E}{V^2} \tag{4a,b}$$

For the undirected graph shown above, eqs. (4) produces $<a_i> = 14/6 = 2.333$, and *Conn* = 14/36 = 0.389 (or 38.9%). The directed graph is less connected that the undirected graph with the same number of vertices and edges, as can be seen from the values obtained, $<a_i> = 1.833$ and *Conn* = 0.306, respecitively.
When dealing with undirected graphs, connectedness is frequently defined slightly differently as $Conn' = 2E/V(V\text{-}1)$. Here, V(V-1)/2 is the number of edges in the maximally connected graph (*complete graph*) having the same number of vertices..

Connectedness is therefore a measure for the **relative** graph connectivity defined within the 0 to 1 range (or within the 0-100% range, after multiplying by 100). Formula (4b) defines graph connectedness in a more general manner, taking into account also the potential availability of non-zero diagonal adjacency matrix entries, $a_{ii} = 1$. The total number of matrix entries in this case is V^2, not $2E/V(V\text{-}1)$. A non-zero diagonal element of adjacency matrix stands for a *loop*, which is an edge emanating from and ending in the same vertex. A loop represents self-interaction of the species described by the network nodes. Such are, for example, protein dimers in protein-protein networks, cannibalistic species in ecological food webs, and others.

5.2.3. Cluster Coefficient and Extended Connectivity

The vertex degree a_i, which counts the nearest neighbors of a vertex i, is not the only *local* connectivity descriptor. More detailed information on the vertex neighborhood is contained in the *cluster coefficient*, c_i. It is defined as the ratio of the number of edges E_i between the first neighbors of the vertex i, and the respective number of edges, $E_i(\max) = a_i(a_i\text{-}1)/2$, in the complete graph that can be formed by the nearest neighbors of this vertex:

$$c_i = \frac{2E_i}{a_i(a_i - 1)} \tag{5}$$

Applying eq (5) to the nondirected graph **1** shown in the foregoing, one obtains for the cluster coefficients the values $c_5 = c_6 = 0$, $c_4 = 1/3$, $c_1 = 2/3$, and $c_2 = c_3 = 1$. In the corresponding directed graph **2**, the cluster coefficient of vertex 4 goes down to zero.
More detailed description of graph connectivity takes into account the second and further neighborhoods. This can be done both locally and globally. The *second cluster coefficient* c_i' counts the edges between the second neighbors of vertex i, and again compares that count to the number of edges in the complete graph that could be formed by all second neighbors. Globally, the layers of second, third, etc., neighbors are taken into account in calculating the graph n^{th}–order extended *connectivity*,[39] ${}^nEC$. The calculation is performed by an iterative procedure, which at each step recalculates the vertex degree of each vertex as the sum of vertex degrees of its first neighbors, as obtained in the previous iteration:

$${}^nEC = \sum_{i=1}^{V} {}^n a_i = \sum_{i=1}^{V} \sum_{j\,adj\,i} {}^{n-1} a_j \tag{6}$$

One may thus form a vector of the extended connectivities of increasing order, $\{EC\} \equiv \{{}^0EC, {}^1EC, {}^2EC, \ldots\}$, the zero-order term in which is the total graph adjacency, defined by eq (2b). Illustration of the iterative calculation of the first several ${}^kEC$ – terms of graph G is shown in Fig. 3.

5.2.4. Graph Distances

In subsection 5.2.1, a *path* in the graph was defined as a sequence of adjacent edges between two vertices without traversing any intermediate vertex twice. The *distance* d_{ij} between vertices *i* and *j* is the shortest path between them. The *distance matrix* $\boldsymbol{D(G)}$ of graph *G* is a square *V* x *V* matrix, which for undirected graphs is symmetrical with respect to the main diagonal. The sum over the matrix row entries is termed *vertex distance degree* of simply *vertex distance*, d_i. The sum over all distance matrix entries is called *graph distance*, *D*:

$$d_i = \sum_{j=1}^{V} d_{ij}; \qquad D(G) = \sum_{i=1}^{V}\sum_{j=1}^{V} d_{ij} = \sum_{i=1}^{V} d_i \tag{7a,b}$$

The average vertex distance (degree) $<d_i>$ and average graph distance $<d>$ (called also *graph radius* or *average path length* or *average degree of vertex-vertex separation*) are also defined:

$$< d_i >= \frac{D}{V}; \qquad < d >= \frac{D}{V(V-1)} \tag{8a,b}$$

Examples illustrating distance matrix and derived descriptors are shown in Fig. 4.

For graphs having loops, the denominator of eq. (8b) changes to V^2 to include the diagonal elements of the distance matrix. *Distance degree distribution* $\{d_i\} \equiv \{d_1, d_2, \ldots, d_V\}$, and *distance magnitude distribution* $\{d\} \equiv \{n_1, n_2, \ldots, n_V\}$ are also defined from the distance matrix, where n_i is the frequency of occurrence of distance with magnitude *i*. *Vertex eccentricity*, e_i, is the maximum distance between vertex *i* and any of the remaining graph vertices. The largest vertex eccentricity is termed *graph diameter.* The vertex(es) with minimum eccentricity is defined as *graph center.*[36] An extended graph center definition[40,41] assumes the minimum eccentricity as a first criterion in a hierarchical series of criteria, which also includes the conditions for the minimum distance degree, and the minimum *distance degree sequence*, *DDS*. The latter is an ascending sequence of the distance magnitudes $1^{n1}\ 2^{n2}\ 3^{n3}\ldots\ (d_{max})^{nmax}$, with each distance frequency *ni* as an exponent. An iterative vertex/edge centricity algorithm IVEC has been developed for the cases when the three hierarchical conditions do not suffice.[42]

The distance degree distributions of graphs **1** and **2** are those given in the d_i columns of the matrices, whereas the distance magnitude distributions of the two graphs are $\{d(G)\} \equiv \{14, 10, 6\}$ and $\{d(DG)\} \equiv \{11, 7, 3\}$, respectively. The vertex eccentricities in graph **1** are $e = 2$ for vertices 4 and 5, and $e = 3$ for the other four vertices. This specifies vertices 4 and 5 as graph centers according to the classical definition of Harary[36], and determines the graph diameter to be equal to 3. The extended graph center definition eliminates vertex 5, due to its larger distance degree (8 vs. 6), and leaves vertex 4 as a single graph center.

Several remarks should be made here related to the distances in directed graphs. Strictly speaking, directed graphs like graph **2** are disconnected, due to the lack of paths between some pairs of vertices, like the missing paths from vertex 6 to all other vertices. The distance between such pairs of vertices is equal to infinity, which makes the calculation

of the total distance in directed graphs impossible. For practical purposes, one might discard such matrix entries as done in $\boldsymbol{D}(\mathbf{2})$ above. However, as pointed out by Neuman *et al*,[43] the distance estimates produced in that way could be totally misleading. Indeed, in comparing the distance estimates for graphs **1** and **2**, e. g., $\langle d(\mathbf{2})\rangle = 1.62 < \langle d(\mathbf{1})\rangle = 1.73$, one may come to the wrong conclusion that the vertices in the directed graph **2** are closer to each other than those in the parent graph **1**. One way toward resolving these difficulties will be shown in Section 5. Another approach to the partial disconnectedness of directed graphs was proposed by Newman *et al*,[43] who introduced the notion of strongly connected, as well as in- and out-component. A *strongly connected component* of a directed graph is a subgraph all vertices in which are connected by a finite path. The *out-component* contains vertices that can reach the strongly connected component but cannot be reached by any vertex of the strongly connected component. Conversely, the *in-component* contains all the vertices that cannot reach the vertices of the strongly connected component but can be reached from them. In the directed graph **2**, used in our examples, one can discern a strongly connected component formed by vertices 1-4, which can reach each other, as can be seen in the distance matrix $\boldsymbol{D}(\mathbf{2})$ above. Vertices 5 and 6 form an in-component; they can be reached from the strongly connected component. The graph lacks an out-component.

Another feature of directed graphs is that the distance degrees d_i defined by eq. (7a) as sums over the matrix row entries are in fact *distance out-degrees*, $d_i(out)$. The *distance in-degrees*, $d_i(in)$, which are obtained as sums over the distance matrix columns

$$d_i(in) = \sum_{i=1}^{V} d_{ij} \tag{7c}$$

are no more the same with their *out*-counterpart, because the directionality of graph arcs destroys the symmetry of the matrix. One may illustrate this point by comparing the two distributions for graph **2**: $\{d_i(\mathbf{2}, out)\} \equiv \{9,9,8,7,1,0\}$ and $\{d_i(\mathbf{2}, in)\} \equiv \{12, 7, 4, 4, 4, 3\}$. Indeed, the total number of *in*- and *out*-distances in a directed graph must be equal. Vertices with large distance *out*-degrees may be of interest in the network analysis as important input nodes, whereas those with large distance *in*-degrees characterize essential output nodes.

Graph centers cannot be rigorously defined in directed graphs containing pairs of vertices with infinite distance between them. However, eliminating such vertices as potential graph centers, one may assess the remaining vertices with the same three criteria discussed above. In- and -out distances may define in principle different vertices as graph centers. In the example with the directed graph **2**, vertex 4 is classified as *out*-center by its minimum out-eccentricity value e_4 (out) = 2 = min (vertices 5 and 6 are excluded from the competition of distance out-degrees). There is no competition for the *in*-center, which is in vertex 6, the only vertex that can be reached by all other vertices (all other vertices are excluded).

5.2.5. Weighted Graphs

An essential generalization of the notion of graph, going beyond topology, enables the application of graph theory to every aspect of cellular networks. One may ascribe different vertex and edge weights, w_{ii} and w_{ij}, to match essential parameters of network species and their interactions. Vertex weights might characterize the level of expression of network species, as measured by mass-spectra, microarrays, HPLC, 2-D gel chromatography, and other methods. The edge weights in metabolic networks might characterize the enzymes expression. An edge weight in networks build of protein complexes denotes the number of proteins two complexes share. Other applications of weighted graphs exist or might be anticipated.

An edge or vertex *weight* could be any nonnegative natural number. (Weights having both positive and negative values has to be renormalized in order to enable using eqs. (9-12). Weights can also be integers, as is the case with multigraphs, in which more than one edge connects some pairs of vertices. Another example is molecular networks, the different chemical nature of the atoms in which is sometimes labeled with vertex weights showing the number of their valence electrons. The *weighted adjacency matrix*, $\boldsymbol{WA}(G)$, has the edge weights w_{ij} as nondiagonal elements, and the vertex weights as diagonal elements, w_{ii}. All graph-invariants derived from the adjacency matrix of a directed or nondirected simple graph can be redefined for a weighted graph. Included here are the *weighted vertex degree*, w_i, and the corresponding *weighted vertex degree distribution*,$\{w_{max}, \ldots, w_{min}\}$,*weighted adjacency*, $WA(G)$, the *average weighted vertex degree*, $<w_i>$, the *weighted connectedness*, *WConn*, the *weighted cluster coefficient*, wc_i, and the *weighted extended connectivity of order k*, ${}^{k}WEC$:

$$w_i = \sum_{j=1}^{V} w_{ij} \; ; WA(G) = \sum_{i=1}^{V}\sum_{j=1}^{V} w_{ij} = \sum_{i=1}^{V} w_i \tag{9a,b}$$

$$< w_i > = \frac{WA}{V}; \qquad WConn = \frac{WA}{V^2} \tag{10a,b}$$

$$wc_i = \frac{\sum_{j\,adj\,i} w_{ij}}{w_i(w_i - 1)} \tag{11}$$

$${}^{n}WEC = \sum_{i=1}^{V} {}^{n}w_i = \sum_{i=1}^{V}\sum_{j\,adj\,i} {}^{n-1}w_j \tag{12}$$

5.3. How to Measure Network Complexity

5.3.1. Careful with Symmetry!

There is a long-term controversy in the literature whether complexity of a structure increases with its connectivity or rather it passes through a maximum and goes down to zero for complete graphs. This is illustrated in Fig. 5 with an example taken from Gell-

Mann's book[44] "*The Quark and the Jaguar*". The example includes two graphs with eight vertices; the first one is totally disconnected, whereas the second one is totally connected (complete) graph. It is argued that the two graphs are equally complex. The arguments in favor of this conclusion are based on the binomial distribution of vertex degrees in random graphs (Fig. 5). Additional arguments in favor of such views come from Shannon's information theory.[1] According to it, the *entropy of information* $H(\alpha)$ in describing a message of N symbols, distributed according to some equivalence criterion α into k groups of $N_1, N_2, \ldots, N_k$ symbols, is calculated according to the formula:

$$H(\alpha) = -\sum_{i=1}^{k} p_i \log_2 p_i = -\sum_{i=1}^{k} \frac{N_i}{N} \log_2 \frac{N_i}{N} \quad \text{bits/symbol} \qquad (13)$$

where the ratio $N_i / N = p_i$ defines the probability of occurrence of the symbols of the i^{th} group.

In using equation (13) to characterize networks or graphs, it is the vertices that most frequently play the role of symbols or system elements. When the criterion of equivalence α is based on the orbits of the automorphism group of the graph, all vertices of the totally disconnected graph belong to a single orbit, and the same is true for the vertices in the complete graph. Eq. (13) then shows that the information index $I(\alpha) = 0$ for both graphs. The same result is obtained when the partitioning of the graph vertices into groups is based on the equality of their vertex degrees, all of which are zeros in the totally disconnected graph, and all of which are of degree N-1 in the complete graph.

The logic of the above arguments seems flawless. Yet, our intuition tells us that the complete graph is more complex that the totally disconnected graph. There is a hidden weak point in the manner the Shannon theory is applied, namely how symmetry is used to partition the vertices into groups. One should take into account that symmetry is a simplifying factor, but not a complexifying one. A measure of structural or topological complexity must not be based on symmetry. The use of symmetry is justified only in defining compositional complexity, which is based on equivalence and diversity of the elements of the system studied.

5.3.2. Can Shannon's Information Content Measure Topological Complexity?

A different approach to characterizing structures by Shannon's theory was proposed in 1977 by Bonchev and Trinajstić in a study on molecular branching as a basic topological feature of molecules.[15] The approach was later generalized by constructing a finite probability scheme for a graph.[16] Let the graph is represented by some kind of elements (vertices, edges, distances, cliques, etc.); let also assign a certain weight (value, magnitude) w_i to each of the N elements. Define the probability for a randomly chosen element i to have the weight w_i as $p_i = w_i / \Sigma w_i$, with $\Sigma w_i = w$, and $\Sigma p_i = 1$. The probability scheme thus constructed

Element	$1, 2, \ldots, N$
Weight	$w_1, w_2, \ldots, w_N$
Probability	$p_1, p_2, \ldots, p_N$

enables defining a series of information indices, $I(w)$, with Shannon's equation (13).

Considering the simplest graph elements, the vertices, and assuming the weights assigned to each vertex to be the corresponding vertex degrees, one easily distinguishes the null complexity of the totally disconnected graph from the high complexity of the complete graph. The probability for a randomly chosen vertex i in the complete graph of V vertices to have a certain degree a_i is $p_i = a_i / A = 1 / V$, wherefrom eq (13) yields for the Shannon entropy of the vertex degree distribution the nonzero value of $\log_2 V$.

Our preceding studies[17, 45-47] have shown that a better complexity measure of graphs and networks is the vertex degree magnitude-based information content, I_{vd}. Shannon defines information as the reduced entropy of the system relative to the maximum entropy that can exist in a system with the same number of elements:

$$I = H_{\max} - H \tag{14}$$

The Shannon entropy of a graph with a total weight W and vertex weights w_i is given by a formula derived from eq (13):

$$H(W) = W \log_2 W - \sum_{i=1}^{V} w_i \log_2 w_i \tag{15}$$

The maximum entropy is obtained when all $w_i = 1$:

$$H_{\max} = W \log_2 W \tag{16}$$

From eqs. (14-16), substituting also $W = A$ and $w_i = a_i$, one obtains the equation for the information content of the vertex degree distribution of a graph, I_{vd}:

$$I_{vd} = \sum_{i=1}^{V} a_i \log_2 a_i \tag{17}$$

The analysis has shown that the I_{vd} index satisfies the criteria for a complexity measure and can be recommended for assessments of network complexity.[17, 45-47] It increases with the connectivity and other complexity factors, such as the number of branches, cycles, cliques, etc., as shown in the series of graphs in Figure 7. The increase in the number of branches increases the complexity index, as seen in the sequences of graphs **3** → **4** → **5**, **6** → **7** → **8**, **9** → **10**, and **12** → **13**. The number of cycles is a considerably stronger complexity factor, as demonstrated in the sequence of graphs with one to five cycles: **6** → **9** → **12** → **14** → **15**.

5.3.3. Global, Average, and Normalized Complexity

A variety of graph-invariants have been examined as measures of topological complexity.[48-50] Since they are directly applicable to networks, we shall review some of the most promising ones, systematizing them in a scheme discussed below.

A series of connectivity descriptors was introduced in section 5.2.2. Total adjacency A is the count of all pairwise neighborhood relationships, $a_{ij} = 1$, each of which denotes a link directed from vertex i to vertex j. Total adjacency is thus equal to the total number of directed edges in the graph. In nondirected graphs, one usually equalizes total adjacency to the doubled number of edges, $A = 2E$. Each nondirected edge $\{ij\}$ in these graphs is in fact an abbreviated notation for two directed edges, one from i to j, and the second one from j to i, respectively. One might then abandon the tradition, and use the symbol E for the total number of (directed, in- and out-) edges in both directed and nondirected graphs, i. e., to use E for the total number of nonzero adjacency matrix entries a_{ij}. We may summarize this analysis by interpreting the redefined total adjacency A as a first level topological complexity measure, and term it graph (or network) *global edge complexity*, E_g.

$$A = \sum_{i=1}^{V} \sum_{j=1}^{V} a_{ij} = \sum_{i=1}^{V} a_i = E_g \tag{18}$$

A similar reinterpretation may be made to the average vertex degree $<a_i>$, and connectedness, *Conn*, introduced by eq (4b). One may call the average vertex degree thus defined *average edge complexity*, E_a, the averaging being defined per vertex. On its turn, connectedness can be regarded as *normalized edge complexity*, E_n, because it is redefined as the ratio of the global edge complexity $E_g = A = E$ and the number of edges in the complete graph with loops at each of its vertices, $E(K_V)$:

$$< a_i > = \frac{A}{V} = \frac{E_g}{V} = E_a \; ; \qquad Conn = \frac{A}{V^2} = \frac{E_g}{V^2} = E_n \tag{19a,b}$$

When the graph contains no loops, the denominator of eq (19b) may be replaced by the V(V-1), eliminating thus the potential contributions from the adjacency matrix diagonal elements of the complete graph.

We have thus presented three individually introduced connectivity descriptors, as three versions of the simplest topological complexity measure: the global, average, and normalized edge complexity. We shall use this triple scheme in presenting other, more sophisticated measures of network complexity. Such more advanced complexity indices are needed because connectedness (the relative edge complexity) is a descriptor that counts only the total number of vertex interconnections, but does not account for the specific way these connections occur. At the same connectedness two networks could differ in their complexity by orders of magnitude. It may be anticipated that the global measures will be of major use in characterizing pathways and small networks, whereas the large networks will be better assessed by the average and relative complexity measures.

5.3.4. The Subgraph Count, *SC*, and Its Components

What would be the next step in the search for more adequate network complexity measures? We started in the preceding subsection with counting the simple subgraphs, the edges, and called this descriptor edge complexity. It seems logical to continue with counting the subgraphs containing two edges. The importance of the two-bonds molecular fragments for the properties of chemical compounds has been early understood, and the total number of these fragments is known in chemical theory as Platt's index.[51] Bertz used this index as a measure of molecular complexity,[17] calling the two-edge fragments "connections". He also constructed an information complexity measure proceeding from the distribution of the two-edge subgraphs into equivalence groups.[18] The Platt index is considerably better complexity measure than the number of edges. At the same number of edges the Platt index increases rapidly with the presence of complexifying factors like branches and cycles.

Such an example is shown in Figure 8, in which graph **1** having two cycles is compared to the path graph **16** having the same number of seven edges. The number of two-edge subgraphs is denoted as ${}^2SC$, meaning 2nd-order subgraph count (*vide infra*). The corresponding average and relative substructure counts of 2nd-order are also shown.

The two graphs differ considerably by their complexity, because the path graph **16** lacks any complexifying structural features, whereas graph **1** incorporates two cycles. Connectedness, *Conn*, does not reflect to a sufficient degree this difference in complexity of the two graphs (*Conn*(**1**) : *Conn*(**16**) = 1.9), whereas the normalized two-edge complexity ${}^2SC_n$ of graph **1** is shown to be much higher than that of **16** (0.5 : 0.036 = 13.9).

In calculating the ${}^2SC_n$ values:

$$ {}^2SC_n = \frac{{}^2SC}{{}^2SC(K_V)} \tag{20} $$

we made use of the formula derived[52] for the 2nd-order subgraph count of the complete graph K_V:

$$ {}^2SC(K_V) = E \times (a_i - 1) = \frac{1}{2}V(V-1)(V-2) \tag{21} $$

The analysis performed in chemical graph theory has shown that the Platt index still fails to mirror some complexity structural patterns, and the search for better measures has continued. A next logical step would be to use the number of three-edge subgraphs, ${}^3SC$. Such an index has been used in chemical graph theory as Gordon-Scantleburry index,[53] however, it has not been tested as a complexity measure. Instead, Bertz and Herndon proposed in 1986 the idea to use the total subgraph count, *SC*, which includes subgraphs of all sizes, including the graph itself, regarded as a proper subgraph.[54] The idea remained unused until the late 1990s, when Bertz[26,27] and Bonchev[9,24,25,28,29] independently and simultaneously developed the approach in detail. Bertz applied the *SC* global index to the synthesis planning in organic chemistry, while the present author derived explicit *SC* formulae for some basic classes of graphs, and the represented the total subgraph count as an ordered set of counts of subgraphs having a given number of edges. The set {*SC*}

begins with the number of vertices V, regarded as null-order index, ${}^0SC$, followed by the number of edges E, as first-order index, ${}^1SC$, the two-edge subgraphs, as the second-order index, ${}^2SC$, etc.:

$$SC = {}^0SC + {}^1SC + {}^2SC + ... + {}^ESC \tag{22a}$$

$$\{SC\} = \{{}^0SC, {}^1SC, {}^2SC, ..., {}^ESC\} \tag{22b}$$

Illustrating the formulas, one obtains for graph **1** the total subgraph count SC = 90, and the set of its null- through seventh-order terms $\{SC\}$ = {6, 7, 12, 20, 22, 16, 6, 1}. The calculations were performed with the program SUBGRAU developed by Rücker and Rücker.[55]

In assessing the complexity of large networks, formulas (22a,b) lead to combinatorial explosion. By this reason, one might recommend using for such purposes only the first-, second-, and third-order subgraph count, whereas the higher orders and the total count could be calculated for pathways and small subnetworks. It is worth mentioning that connectedness (or connectance), which is used almost exclusively in characterizing dynamic networks, appears naturally as the normalized first-order term in the series (22a,b). One might anticipate a broader application of the higher terms, particularly ${}^2SC_n$ and ${}^3SC_n$, due to their much higher sensitivity to the complexifying details of the networks. For the normalizing of these terms one may use the formulas we derived for the three-edge subgraph count ${}^3SC$ of the complete graph K_V, as well as for its components, the counts of triangular, linear, and star type three-edge subgraphs:

$${}^3SC(K_V) = \frac{1}{6}V(V-1)(V-2)(4V-11) \tag{23}$$

$${}^3SC(K_V, triangle) = \frac{1}{6}V(V-1)(V-2) \tag{24}$$

$${}^3SC(K_V, linear) = \frac{1}{2}V(V-1)(V-2)(V-3) \tag{25}$$

$${}^3SC(K_V, star) = \frac{1}{6}V(V-1)(V-2)(V-3) \tag{26}$$

The comparison of the third-order subgraph counts of graphs **1** and **3**, 20 vs. 5, shows again a considerably higher complexity of graph **1** as compared to the assessment based on the graph connectedness (connectance). One may also recommend to use for more detailed characterization of complex networks, the separate counts of the three kinds of three-edge subgraphs – triangles, stars, and linear ones, ${}^3SC_t$, ${}^3SC_s$, and ${}^3SC_l$, which were previously shown to produce high correlations with physicochemical properties.[56]

5.3.5. Overall Connectivity, *OC*

The subgraph count presentation as an ordered set of components with increasing size may be regarded as a part of a more general scheme.[57] The latter defines a certain *overall* graph-invariant X, by the sum over the values this invariant has for each of the subgraphs. Also, the contributions of all subgraphs having k edges are combined in single term, ${}^kX$. An ordered set $\{X\}$ on all k-terms is also constructed, and the initial terms k = 0,1,2,3,…, called null-, first-, second-, etc. order terms, can be independently used to characterize the graph properties.

$$X = \sum_{k=1}^{E} {}^kX; \quad \{X\} = \{{}^0X, {}^1X, {}^2X, ..., {}^EX\} \tag{27}$$

In addition, one can also define the average value of X per vertex, X_a, as well as its normalized value, $0 \le X_n \le 1$:

$$X_a = \frac{X}{V}; \quad {}^kX_a = \frac{{}^kX}{V} \tag{28a,b}$$

$$X_n = \frac{X}{X(K_V)}; \quad {}^kX_n = \frac{{}^kX}{{}^kX(K_V)} \tag{29a,b}$$

The scheme can be further detailed by using within each ${}^kX$ term the counts of subgraphs of different topology, e.g., for three edge subgraphs the counts of triangles, stars, ane linear (or path) graphs. [56]

The simplest graph-invariant that can be incorporated into this scheme is the subgraph count, *SC*, as shown in the foregoing. The next basic candidate is the graph adjacency A, defined by eq (2b). By summing up the adjacencies of all k^{th}-order subgraphs ${}^kG_i$, with k = 0, 1, 2, 3, …, E, one defines[28,29] the *overall connectivity* $OC(G)$ of the graph G:

$$OC(G) = \sum_{k=1}^{E} {}^kOC = \sum_{k=1}^{E} \sum_{i} {}^kA_i ({}^kG_i \subset G) \tag{30a}$$

$$\{OC\} = \{{}^0OC, {}^1OC, {}^2OC, ..., {}^EOC\} \tag{30b}$$

Eqs. (30a,b) yield for graph **1** the overall connectivity value OC = 936, and the set of its 0- to 7-th order terms: $\{OC\}$ = {14, 38, 101, 210, 264, 212, 83, 14}. It should be mentioned that in the first publications defining overall connectivity,[24, 25] the latter was termed *topological complexity* and denoted by *TC*. This name was later changed[28, 29] to overall connectivity to account for the fact that this is not the only measure of topological complexity.

According to the general scheme, the overall connectivity index can also be presented as averaged per vertex, and in a normalized form. To facilitate the calculation of the first-,

second-, and third-order normalized index, eqs. (31-33) were derived, along with eqs. (34-36) for the three different topological shapes of the three-edge subgraphs:

$$^{1}OC(K_V) = V(V-1)^2 \tag{31}$$

$$^{2}OC(K_V) = \frac{3}{2}V(V-1)^2(V-2) \tag{32}$$

$$^{3}OC(K_V) = \frac{1}{6}V(V-1)^2(V-2)(16V-45) \tag{33}$$

$$^{3}OC(K_V, triangle) = \frac{1}{2}V(V-1)^2(V-2) \tag{34}$$

$$^{3}OC(K_V, linear) = 2V(V-1)^2(V-2)(V-3) \tag{35}$$

$$^{3}OC(K_V, star) = \frac{2}{3}V(V-1)^2(V-2)(V-3) \tag{36}$$

The overall topological indices scheme, defined by eqs. (27- 29), has also been applied to other graph invariants, such as the Wiener number[58-60] and the Zagreb indices.[56,61,62] These overall indices have also shown properties of complexity measures.

5.3.6. The Total Walk Count, *TWC*

Rücker and Rücker have proposed[30,31] a similar scheme for assessing the graph complexity by the *total walk count*, TWC. This complexity measure is obtained by counting all walks $^{l}w_i$ of all lengths l, the maximum walk length being limited by the graph size:

$$TWC = \sum_{l=1}^{V-1} {}^{l}WC = \sum_{l=1}^{V-1}\sum_{i} {}^{l}w_i \tag{37}$$

(Scheme 1 here!)

For graph **1**, one finds $TWC = 1154$ {14, 38, 100, 272, 730}. The length-one walks are just the doubled number of edges, since each of the two ends of an edge is used as a walk starting point. There are two types of walks of length two: forward and back along the same edge (1→2→1) and forward along two adjacent edges (1→2→4). Each of these two types then generates two different types of walks of length three, with the third step backside (1→2→1→2; 1→2→4→2) or along a different edge (1→2→1→4; 1→2→4→3) , etc.

The number of walks of length l, is obtained from the l^{th} power of the adjacency matrix. For calculating the normalized ${}^{l}WC_n$ indices, one has to use eq. (38) derived for the respective value in the complete graph with the same number of vertices. One would then find for graph **1**, ${}^{2}WC_n = 0.253$ and ${}^{3}WC_n = 0.133$.

$$ {}^{l}WC(K_V) = V(V-1)^{l} \tag{38} $$

Like the subgraph count and the overall connectivity, the total walk count is an adequate measure of graph complexity, showing patterns of regular increase with the graph size, connectedness, and the basic structure complexifying factors such as the number, size and the kind of interconnectedness of the graph cycles and branches.[31] Figure 9 illustrates these conclusions, providing the same ordering of increasing complexity of graphs **3** to **15** like the one produced in the foregoing by the I_{vd} index.
The complexity measures discussed in Section 3 have all been previously published. In the next Section 4, we report some new developments.

5.4. Combined Complexity Measures Based on the Graph Adjacency and Distance

5.4.1. The *A/D* Index

Networks with high complexity are characterized by both high vertex-vertex connectedness and small vertex-vertex separation (the small-world concept of Watts and Strogatz[63]). Therefore, it seems logical to use both quantities in characterizing network complexity. The ratio $A/D = \langle a_i \rangle / \langle d_i \rangle$ of the total adjacency and the total distance of the graph or, equivalently, the ratio of the average vertex degree $\langle a_i \rangle$ and the average distance degree $\langle d_i \rangle$, may be regarded as a logical approach to such a complexity measure. At a constant number of vertices, the *A/D* index has a minimum value in path graphs, P_V, which are characterized by low connectivity and long distances. In contrast, the *A/D* ratio has a maximum value in the complete graphs, K_V, which are maximally connected and all of their vertices have only a unit distance separation. The classes of star graphs, S_V, and monocyclic graphs, C_V, are of intermediate complexity and their *A/D* indices are between these two extremes.

$$ A/D(P_V) = \frac{2(V-1)}{V(V-1)(V+1)/3} = \frac{6}{V(V+1)} \tag{39} $$

$$ A/D(K_V) = \frac{V(V-1)}{V(V-1)} = 1 \tag{40} $$

$$ A/D(S_V) = \frac{2(V-1)}{2(V-1)^2} = \frac{1}{V-1} \tag{41} $$

$$A/D(C_V, odd) = \frac{2V}{2V(V^2-1)/8} = \frac{8}{(V^2-1)} \tag{42a}$$

$$A/D(C_V, even) = \frac{2V}{V^3/4} = \frac{8}{V^2} \tag{42b}$$

As shown in eq. (40), the *A/D* index of the complete graph is equal to a unity; therefore, all graphs have their *A/D* values within the 0 to 1 range. Like all normalized complexity indices this index decreases rapidly with the graph size for path graphs, monocyclic graphs, and other weakly connected graphs, the distance in which dominates strongly over adjacency. Some degeneracy of the index (having two or more nonisomorphic graphs with the same *A/D* ratio) should be expected, because both the total adjacency *A* and the total distance *D* are degenerate. What might be a more serious problem is the insensitivity to some more subtle topological features of branching and cyclicity, which sometimes produces incorrect assessments of graph complexity (See Table 1, and the examples in the next subsection). Yet, the fine details of topological structure might be inessential when dealing with large networks, for which the A/D index could prove to be a sufficiently accurate measure of structural complexity. For smaller subnetworks and particularly pathways, perhaps a better recommendation would be to make use of the new structural index presented in Subsection 4.2.

5.4.2. The Complexity Index *B*

The ratio $b_i = a_i / d_i$ of the vertex degree a_i and its distance degree d_i is a local invariant with interesting centric properties. It is ≤ 1, the equality occurring for the central vertex in the star graphs, as well as for every vertex in the complete graph. The sum over the b_i values of all graph vertices may be expected to behave similarly to the *A/D* ratio, with less degeneracy, and more sensitivity to local topology. We define this sum as a new complexity index *B*:

$$B = \sum_{i=1}^{V} \frac{a_i}{d_i} \tag{43}$$

Several equations derived for the b_i and *B* indices shed some light on the properties of these complexity descriptors. In complete graphs, K_V, in which $a_i = d_i = V-1$, and $b_i = 1$ for every vertex, the *B* index is simply equal to the number of vertices *V*:

$$B(K_V) = V \tag{44}$$

In star graphs, S_V, in which the central vertex *c* is of degree *V*-1, and all other vertices are terminal (*t*) with degree 1, one obtains

$$b_t = \frac{1}{2V-3}; \qquad b_c = 1; \qquad B(S_V) = \frac{3V-4}{2V-3} \tag{45a,b,c}$$

In (mono)cyclic graphs, C_V, all vertices have degree two, and have the same distance degree. The expression for the latter differs slightly for the odd- and even-membered cycles:

$$C_V(odd): \quad b = \frac{8}{V^2 - 1}; \qquad B = \frac{8V}{V^2 - 1} \tag{46a}$$

$$C_V(even): \quad b = \frac{8}{V^2}; \qquad B = \frac{8V}{V^2} = \frac{8}{V} \tag{46b}$$

The B index values begin at $B = 3$ for the odd-membered cycles and at $B = 2$ for the even-membered cycles, and gradually decrease with the cycle size to the zero limit at $V \to \infty$.

In the path graphs, P_V, the two terminal vertices are of degree 1 and all others are of degree two. The formulas for the local b_i indices depend on the position $i = 1, 2, 3, \ldots, V$ of the vertex, counting from the end of the chain. Different equation is obtained only for the central one or two vertices c:

$$b_i = \frac{2a_i}{V^2 - (2i-1)V + 2i(i-1)} \tag{47a}$$

$$b_c(odd) = \frac{8}{V^2 - 1}; \quad b_c(even) = \frac{8}{V^2} \tag{47b}$$

No closed form equation can be obtained for the B index of path graphs. However, the presence of the V^2 term in the denominator of the local b_i and b_c indices shows that at large path length they, as well as well the B index, will tend to zero considerably faster than the respective indices for the monocyclic graphs, which decrease with V only linearly.

The testing of the new complexity measure with graphs **3** – **15**, used in Section 3 to demonstrate the behavior of other complexity measures, has shown a perfect match with the ordering produced by the subgraph count, overall connectivity, total walk count and the information on the vertex degree distribution (Figure 10).

The *A/D* index also captured the basic complexity features in this series of graphs to increase with the number of branches and cycles. However, it is less sensitive to subtle details of graph topology, which resulted in three inverse orderings and three degeneracies.

B ordering: **3**(1.105) → **4**(1.294) → **5**(1.571) → **6**(1.667) → **7**(1.677) → **8**(1.783) → **9**(2.200) → **10**(2.211) → **11**(2.410) → **12**(2.867) → **13**(2.943) → **14**(4.200) → **15**(5.000)

A/D ordering: **3**(0.200) → **4**(0.222) → **5**(0.250) → **7**(0.313) = **8**(0.313) → **6**(0.333) → **10**(0.400) → **9**(0.429) = **11**(0.429) → **12**(0.538) = **13**(0.538) → **14**(0.818) → **15**(1.000)

Additional comparisons between the new *A/D* and *B* indices and the four selected known complexity measures are shown in Table 1 for the 13 six-vertex graphs from Figure 11. Once again, the *B* index captures the complexity features of the graphs examined much better than the *A/D* ratio. The A/D index not only shows high degeneracy but in the degenerate quartet and triplet of graphs it produces the same complexity estimate for graphs that all other five indices distinguish drastically, e.g., **18** and **20, 21** and **24**, and others. The *B* index generates the same ordering as the total walk count *TWC*, and has minimal number of reorderings (denoted by asterisks in Table 1) with the subgraph count *SC*, the information index for the vertex degree distribution I_{vd}, and the overall connectivity index *OC*, the latter four indices not producing identical orderings as well. The B index has also a single degeneracy, slightly worse than *OC* and TWC with no degeneracy, and better than *SC* with two, and I_{vd} with even six degenerate values. All this characterizes the index B introduced here as a convenient measure of graph complexity, a measure that shows similar behavior to other well established and sensitive complexity measures, and does not require substantial computational time.

5.5. Vertex Accessibility and Complexity of Directed Graphs

In subsection **5.2.4** we have discussed the misleading results that are obtained for the graph radius (the average path length or the average graph distance) in directed graphs when one simply neglects the infinite distances between the pairs of vertices for which no path exists, and averages the remaining distances. Such calculations would produce the false impression that the radius of directed graphs is smaller than that of the parent undirected graph. A correcting procedure that restores the normal distance ratios between the parent undirected graph and the directed graphs generated from it was recently described.[45] It introduces a parameter called *vertex accessibility*, *Acc*(*DG*), which accounts for the degree to which the vertices in directed graphs are mutually accessible via finite paths. The vertex accessibility of a directed graph *DG* is defined as the ratio of the number of finite distances in the directed graph, $N_d(DG)$, and the total number of distances in the parent undirected graph $N_d(G)$:

$$Acc(DG) = \frac{N_d(DG)}{N_d(G)} \tag{48}$$

In eq (48), $N_d(G) = V^2$ (the squared total number of vertices *V*) in the general case of connected undirected graphs with loops. In that case, $N_d(DG)$ includes also the number of loops, as given with all $d_{ii} = 1$ appearing in the main diagonal of the distance matrix. If no loops can in principle exist in a certain type of networks, then $N_d(G) = V(V-1)$ should be used.

Eq. (48) enables obtaining a more realistic estimate of the average path length $\langle d \rangle$ in a directed graph. Dividing $\langle d \rangle = D/N_d$, by the vertex accessibility, one normalizes this quantity to the case of complete vertex accessibility. The *adjusted average distance* (*adjusted average path length*), *AD*(*DG*):

$$AD(DG) = \frac{< d >}{Acc} = \frac{D \times V^2}{N_d{}^2} \tag{49}$$

thus defined, is larger than the average distance in the parent undirected graph, and can be used for comparisons of the average degree of separation in directed graphs. As in eq. (48), for DGs without loops V^2 can be replaced by $V(V\text{-}1)$. The calculation made for the vertex accessibility of directed graph **2** (see the distance matrix of this graph in subsection **5.2.4**) produces $ACC(\mathbf{2}) = 21/(6\text{x}5) = 0.7$. From here, with eq. (48) one obtains for the adjusted average distance of this graph, $AD(\mathbf{2}) = 1.62/0.7 = 2.31$. Thus, the unrealistic value of 1.62, after the adjustment turned from smaller to considerably larger than the corresponding value of 1.73 for the parent undirected graph 1.

Vertex accessibility can also be used to define a more realistic measure of the connectedness of directed graphs. The new measure might be termed *accessible connectedness*, $AConn(DG)$:

$$AConn(DG) = Conn(G) \times Acc(DG) = Conn(G) \times \frac{N_d(DG)}{N_d(G)} \tag{50}$$

Illustrating eq. (50), the calculation for the directed graph **2** results in $AConn(\mathbf{2}) = 0.214$, down from the unadjusted value of $Conn(\mathbf{2}) = 0.306$ calculated in subsection **5.2.2**, a value that was unrealistically close to that of the parent undirected graph $Conn(\mathbf{1}) = 0.389$.

Similar adjustment may be made to the *A/D* index of directed graph. Substituting the misleading distance *D* by its adjusted counterpart *AD*, one defines the *A/AD* complexity measure of directed graphs.

Some classes of directed graphs are of interest, because of the special relations existing for their vertex accessibility and the adjusted indices derived from it. Such is the special class in which all edges are directed and their direction is the same (all linear or clockwise, etc.). It can be easily shown that for monocyclic and complete graphs of this class, there is a complete accessibility of all vertices, at the cost of considerably larger average path length than that of the parent undirected graph. Thus, the directed graph DC_6 has a total distance of 90, a vertex distance of 15, and an average distance of 3, whereas its parent undirected graph C_6 has a total distance of 54, a vertex distance of 9, and an average distance of only 1.8. The directed graph DK_5 has a total distance of 30, a vertex distance of 6, and an average distance of 1.5, as compared to the parent complete graph K_5 having a total distance of 20, a vertex distance of 4, and an average distance of 1.

The directed path graph and star graph shown in Fig. 12 do not have complete vertex accessibility. The actual accessibility, the adjusted average distance, and the adjusted connectedness can be assessed by the following formulae:

$$Acc(DP_V) = \frac{1}{2}; \quad AD(DP_V) = \frac{2(V+1)}{3}; \quad AConn(DP_V) = \frac{1}{2V} \tag{51a,b,c}$$

$$Acc(DS_V, odd) = \frac{V+3}{4V}; \quad Acc(DS_V, even) = \frac{V_2 + 2V - 4}{4V(V-1)} \qquad (52a,b)$$

$$AD(DS_V, odd) = \frac{8V(V+1)}{(V+3)^2}; \quad AD(DS_V, even) = \frac{8V(V-1)(V^2-2)}{(V^2+2V-4)^2} \qquad (53a,b)$$

$$AConn(DS_V, odd) = \frac{V+3}{4V^2}; \quad AConn(DS_V, even) = \frac{V^2+2V-4}{4V^2(V-1)} \qquad (54a,b)$$

5.6. Complexity Estimates of Biological and Ecological Networks

Networks are universal means for analyzing systems in their entirety, and for capturing the systems complexity patterns.[64] Not surprisingly, after the revolution in network theory started[65]in 1999, and the focus has shifted from random networks to dynamic evolutionary ones,[66] up to a half of all working papers of the Santa Fe Institute of Complexity have been devoted to networks.[67] The physical nature of the network nodes and their interactions is inessential in this analysis. In biological networks nodes can represent proteins[68-71] or protein complexes,[72] genes,[73-75] metabolites,[76-78] neurons,[79] etc. The type of "interaction" that connects two nodes in the network in an edge or arc could also vary from chemical binding to regulatory effects to signal transduction to nerve impulse. There are also networks in which there is no real interaction but the edge may stand, for example, for the presence of the same species (proteins or genes) in different complexes. In food webs, the nodes represent different kind of biological species, while the type of interaction is "who is eating whom". However different systems the networks may represent, they all have common features and share common structural patterns based on the connectivity of their constituents. Complexity measures make possible the characterization of these common network features in a general quantitative scale, providing thus the means for comparisons and quantitative evolutionary models.

5.6.1. Networks of protein complexes

Proteins tend to associate with each other forming complexes. The size of these complexes may vary within a rather broad range. Fig. 13 presents the network of protein complexes taking part in the RNA metabolism of Saccharomyces Cerevisiae (data taken from Gavin *et al* [72].) The 28 complexes contain 692 proteins, which amounts in average to almost 25 proteins in a complex, the actual sizes ranging from 2 to 133 complexes. The complexes are denoted by sequential numbers as given in the Supplementary Table 3 of the data source[72]. Each edge in Fig. 13 stands for sharing proteins between the corresponding two complexes. The exact number of shared proteins is not shown as edge weights, due to the graph complexity. In the majority of cases the pairs of complexes share only one protein. In four cases, the number of shared proteins is between ten and fifteen. The calculations of the complexity measures of this weighted undirected graph are also performed for the basic topology of the parent non-weighted graph.

The graph actually shows the giant component (a term used to denote the graph component that incorporates the majority of vertices) of the network, the latter also containing three complexes that not share proteins with other complexes. The giant component is highly connected with a 106 non-weighted edges or basic adjacency of 212. This leads to average basic vertex degree of 8.48, and connectedness of 0.353. The corresponding values based on the edge weights are: weighted vertex adjacency of 1124, average weighted vertex degree of 44.96, and weighted connectedness of 1.87. This high connectedness evidences for the high stability against attacks or mutations, and indicates the importance of the RNA metabolism for the cell survival. High adjacency/ connectedness values are obtained also for the networks of protein complexes responsible for transcription/DNA maintenance/chromatin structure, and for protein synthesis and turnover (Table 2). The comparison of the connectivity descriptors in Table 2 also allows concluding that the biological functions of signaling, cell cycle, and cell polarity and structure are more vulnerable against such attacks. Similar conclusions can be drawn from Table 3, which presents the values of the more recent complexity measures calculated for the weighted graph (not shown) derivative of graph given in Fig. 13. The six measures included: the two normalized subgraph count descriptors, ${}^{2}SC$ and ${}^{3}SC$, the two normalized overall connectivity indices, ${}^{1}OC$ and ${}^{2}OC$, the normalized information index for the vertex degree distribution, $I_{vd,n}$, and the newly developed *A/D* index, order the protein functional groups in a similar manner. They all single-out the functional group of protein complexes involved in the RNA metabolism as the most complex one, the next two places being occupied alternatively by the group controlling transcription, DNA maintenance, and chromatin structure, and the one of protein synthesis. The A/D index reproduces with a single exception the same ordering, and thus demonstrated its potential as complexity measure. It should be mentioned, that all our calculations were performed with data[72] that comprise about a quarter of all yeast proteins. Accounting for all protein complexes will indeed change the complexity measures values. One may anticipate that the availability of the complete set of data will enable the complexity estimates of performance stability of the biological functions related to cell cycle, cell polarity and structure, and signaling. One may also expect the major conclusion about the three groups of biological function that are best protected against any kind of damage to be confirmed by such more complete analysis.

5.6.2. Food Webs

Food webs are presented by directed graphs, because the interaction between the species is in the great majority of cases unidirectional (the pray cannot eat the predator). Other examples of directed networks are gene regulatory networks and cellular signal transduction networks. It has been shown[64, 81, 82] that the more complex directed networks have a specific structure. It includes in- and out-components, a strongly connected component and a tube (Fig. 14). The nodes in the *strongly connected component* are accessible to each other. These nodes have also incoming edges (arcs) originating from the *out-component*, and outgoing arcs directed to the *in-component*. Vertices from the in-component can also be directly connected to vertices of the out-component thus forming a *tube*.

As shown in the St. Martin island wood web,[83] analyzed below (Fig. 15), this specific hierarchical structure of directed networks is not always possible. The web incorporates 42 trophic species with a total of 205 interactions. The network of this ecological system is rather complex. Nevertheless, it does not have even a triplet of mutually accessible vertices, which to form a strongly connected component. What appears as more essential and always preserved in such networks is their hierarchical structure, based on the principle of downstream interactions. In Fig. 15, the St. Martin island wood web is presented schematically by two different directed graphs. The first graph is composed of six ordered layers A to F. The species of each layer can eat all downstream species, and in a few cases another species of the same layer. This graph shows explicitly only the interactions between the pairs of neighboring layers. The total amount of interactions between all pairs of layers is shown as edge weights in the second graph in Fig. 15, the vertices in which depict the six layers of web species.

The connectivity of the St. Martin's island food web can be characterized by the values of the average vertex degree, a_i = 4.88, and that of connectedness (connectance) = 0.119. (Both values are just half of the corresponding values for the parent undirected graph.) The normalized ${}^2SC$ and ${}^3SC$ complexity indices are equal to 0.0673 and 0.0193, respectively. These values are the same for the directed graph and its undirected parent graph. The first three overall connectivity indices, ${}^1OC$, ${}^2OC$, and ${}^3OC$, are calculated in separate in- and out-terms (in: 0.0387, 0.0119, and 0.0037, and out: 0.0328, 0.0093, and 0.0028, respectively). The sums of the pairs of in- and out-terms, are equal to the corresponding parent undirected graph values. Therefore, the calculation of the in- and out-terms makes sense mainly when comparing different directed graphs DG_i originating from the same parent undirected graph G. In the case of DGs obtained from different Gs, one may use for approximate estimates the complexity measures as calculated for the corresponding parent graphs. The normalized information index on the vertex degree distribution also correctly reproduces the lower complexity of the directed graph relative to that of the parent undirected graph ($I_{vd,n}$ (out) = 0.367, $I_{vd,n}$ (in) = 0.388, $I_{vd,n}$ (G) = 0.401).

There is no such correspondence between the distance measures of directed and parent undirected graphs, due to the lack of paths between some pairs of vertices in the DGs. Thus, while the undirected St. Martin graph has 1722 vertex-vertex distances, in the directed graph they are only 446 (205x1, 209x2, 32x3). The total distance calculated from these is 719 vs. 3308 in the undirected graph. Comparing the average distances of the two graphs would be misleading, because it would show the vertices of directed graph to be closer to each other than they are in the undirected graph (1.61 vs. 1.92). The things come back to normal after calculating the accessibility of the DG vertices (eq. 48), Acc = 0.259, wherefrom eq (49) produces the more realistic value of $<d(DG)> = 6.22 > 1.92$. More realistic estimate of the directed graph connectedness may also be obtained by eq (50), accounting for the limited vertex accessibility: $AConn(DG) = 0.031 < Conn(DG) = 0.119$, the latter value being unrealistically close to that of the undirected graph connectedness (0.238). Similar correction might be made for the A/D complexity index introduced in Section 4.1. This index shows a pattern of continuous increase with the increase in the network complexity. However, the value calculated for the directed graph, $A/D(DG) = 205/719 = 0.258$ is larger than that of the undirected graph, A/D(G) = 410/ 3308 = 0.124. The higher complexity of the undirected graph can be correctly assessed by

adjusting the A/D index by multiplying it by the accessibility index (0.258 x 0.259 = 0.067 < 0.124).

The different complexity indices order the food web in a similar manner (Figure 16). The connectedness index cannot distinguish two pairs of food webs (St. Martin Island/Lake Little Rock, Conn = 0.119, and Skipwith Pond/Coachella Valley, Conn = 0.328/0.323), whereas the latter are well discerned by the subgraph count and overall connectivity indices. Conversely, ${}^{2}OC$ and ${}^{3}OC$ cannot well discriminate Ythan Estuary and Canton Creek food webs.

Many studies have shown than a higher connectivity and complexity means a higher network stability.[84, 85] One may thus expect the Skipwith Pond and Coachella Valley food webs to be very stable to attacks and environmental changes. As recently shown,[45] the Skipwith Pond ecosystem could survive even the elimination of half of its best connected trophic species in the food web. The least complex webs examined - those of Ythan Estuary and Canton Creek - may be expected to be more vulnerable. To verify this conclusion, we modeled the specific attack on this web by subsequently eliminating its highest-degree vertices. It was found that after eliminating the first 13 such vertices, which corresponds to a 2-fold decrease in the web connectedness and to a 12-fold decrease in the web complexity as described by the ${}^{2}SC$ and ${}^{1}OC$ indices, the network splits into a large and a small component (Figure 17).

5.8. Overview

In this chapter, we reviewed some of the complexity measures, which were shown in previous publications to be appropriate for assessments of network complexity. A clear distinction was made between the two types of complexity: the compositional and the structural (topological) ones. Four topological complexity measures were presented in detail: the information on the vertex degree distribution, the subgraph count, the overall connectivity, and the walk count. The last three were presented as ordered sequences of terms corresponding to subgraphs with increasing number of edges. Equations were derived for the first several orders of each of the complexity descriptors, which will facilitate their application to large scale networks. In addition, each of these measures was presented in three versions: total (or overall), average, and normalized (within the 0 to 1 range) ones. Two new complexity indices were proposed based on the combined use of the adjacency and distance matrix of the network. These indices unite the intuitive ideas of structural complexity resulting from high connectivity and small vertex separation (the “small world” concept). Important corrections were introduced to the way the total distance and the connectedness of directed graphs are calculated, by accounting for the mutual accessibility of network vertices. The mathematical tools introduced were illustrated with numerous examples, including protein-protein interaction networks and food webs. The authors anticipate a wider use of the presented complexity measures for the characterization of network topology, which usually does not go beyond connectedness (connectance), cluster coefficients, and graph radius.

Despite of the rapid development of complexity theory during the last 20 years, one can still face questions like: “Can we measure complexity, and if we can why?” We hope that this chapter answers explicitly the first question. As for the second one, we would like to remind the words of Lord Kelvin, said 150 years ago: “One cannot describe the Laws of

Nature unless he uses numbers." Are there laws of nature related to complexity of systems? Up to very recently, there was no clear idea how to define complexity as a universal property of systems in nature and technology. The situation changed dramatically after Barabási[65] proposed in 1999 to consider the nonrandom dynamic networks as a universal language to describe complexity and evolution of systems. Life sciences have found in cellular networks (protein, gene, and metabolic ones) their long searched tool to describe the work of the biological machine as a whole. It is believed that the next 10-15 years will be the most important ones in the history of biology and medicine. The theory of network complexity will play an important role during this exciting time.

Acknowledgment. The authors are indebted to Drs. G. Rücker and C. Rücker (Bayreuth) for the use of their computer programs SUBGRATCAU and MOR5AU, and to Dr. J. A. Dunne (Santa Fe) for providing the food webs data. D. Bonchev was supported by NIH grant No. 5-22405.

5.8. References

1. C Shannon and W. Weaver, *Mathematical Theory of Communications*, University of Illinois Press, Urbana, IL, 1949.
2. H Kastler, Ed. *Essays on the Use of Information Theory in Biology,* University of Illinois Press, Urbana, IL, 1953.
3. H Linshitz, The Information Content of a Bacterial Cell. In: H Kastler, Ed. *Essays on the Use of Information Theory in Biology.* University of Illinois Press, Urbana, IL, 1953.
4. N Rashevsky, Life, Information Theory, and Topology, *Bull. Math. Biophys.* 17, 229-235, 1955.
5. E Trucco, A Note on the Information Content of Graphs, *Bull Math. Biophys.* 18, 129-135, 1956.
6. A Mowshowitz, Entropy and the Complexity of Graphs. I. An Index of the Relative Complexity of a Graph, *Bull. Math. Biophys.* 30, 175-204, 1968.
7. D Minoli, Combinatorial Graph Complexity, *Atti. Acad. Naz. Lincei Rend.* 59, 651-661, 1976.
8. AN Kolmogorov, Three Approaches to the Quantitative Definition of Information, Problem'i Peredachi Informatsii (Russ.) 1, 1–7, 1965.
9. D Bonchev, Kolmogorov's Information, Shannon's Entropy, and Topological Complexity of Molecules, *Bulg. Chem. Commun*. 28, 567-582, 1995.
10. D Bonchev, D. Kamenski, and V. Kamenska, Symmetry and Information Content of Chemical Structures, *Bull. Math. Biol*. 38, 119-133, 1976.
11. D Bonchev, and N Trinajstić, Information Theory, Distance Matrix, and Molecular Branching, *J. Chem. Phys.* 67, 4517-4533, 1977.
12. D Bonchev and V Kamenska, Information Theory in Describing the Electronic Structure of Atoms, *Croat. Chem. Acta* 51, 19-27, 1978.
13. D Bonchev, Information Indices for Atoms and Molecules, *MATCH - Commun. Math. Comput. Chem.* 7, 65-113, 1979.
14. D Bonchev, O Mekenyan, and N Trinajstić, Isomer Discrimination by Topological Information Approach, *J. Comput. Chem*. 2, 127-148, 1981.
15. D Bonchev and N Trinajstić, Chemical Information Theory, Structural Aspects. *Intern. J. Quantum Chem. Symp*. 16, 463-480, 1982.
16. D Bonchev, Information-*Theoretic Indices for Characterization of Chemical Structures*. Research Studies Press, Chichester, UK, 1983.
17. D. Bonchev, Shannon's Information and Complexity. In: *Mathematical Chemistry Series*, *Vo*l. 7, *Complexity in Chemistry*, D Bonchev and DH Rouvray, Eds., Taylor & Francis, London, 2003, p 155-187.
18. SH Bertz, The First General Index of Molecular Complexity, *J. Am. Chem. Soc.* 103, 3599-3601, 1981.
19. SH Bertz, The Bond Graph, *J. Chem. Soc. Chem. Commun.* 209, 1981.
20. D Bonchev, The Problems of Computing Molecular Complexity, In: *Computational Chemical Graph Theory*, DH Rouvray, Ed., Nova Publications, New York, 1990, p. 34-67.
21. S Nikolić, N Trinajstić, M Tolić, G Rücker and C Rücker, On Molecular Complexity Indices, In: *Mathematical Chemistry Series*, *Vol*. 7, *Complexity in*

Chemistry, D Bonchev and DH Rouvray, Eds, Taylor & Francis, London, 2003, p 29-89.

22. SH Bertz, A Mathematical Model of Molecular Complexity, In: *Chemical Applications of Topology and Graph Theory*, RB King, Ed., Elsevier, Amsterdam, 1983, p. 206-221.
23. D Bonchev and OE Polansky, On the Topological Complexity of Chemical Systems, In: *Graph Theory and Topology in Chemistry*, RB King and DH Rouvray, Eds, Elsevier, Amsterdam, 1987, p.126-158.
24. D Bonchev and WA Seitz, The Concept of Complexity in Chemistry. In: *Concepts in Chemistry: A Contemporary Challenge*, DH Rouvray, Ed, Wiley, New York, 1997, p. 353-381.
25. D Bonchev, Novel Indices for the Topological Complexity of Molecules, *SAR QSAR Environ. Res.* 7, 23-43, 1997.
26. SH Bertz and TJ Sommer, Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually Simple New Complexity Indices, *Chem. Commun.* 2409-2410, 1997.
27. SH Bertz and WF Wright, The Graph Theory Approach to Synthetic Analysis: Definition and Application of Molecular Complexity and Synthetic Complexity, *Graph Theory Notes New York Acad. Sci.* 35, 32-48, 1998.
28. D Bonchev, Overall Connectivity and Molecular Complexity. In: *Topological Indices and Related Descriptors,* J Devillers and AT Balaban, Eds. Gordon and Breach, Reading, UK, 1999, p. 361-401.
29. D Bonchev, Overall Connectivities /Topological Complexities: A New Powerful Tool for QSPR/QSAR, *J. Chem. Inf. Comput. Sci*. 40, 934-941, 2000.
30. G Rücker and C Rücker, Walk Count, Labyrinthicity and Complexity of Acyclic and Cyclic Graphs and Molecules, *J. Chem. Inf. Comput. Sci.* 40, 99-106, 2000.
31. G Rücker and C Rücker, Substructure, Subgraph and Walk Counts as Measures of the Complexity of Graphs and Molecules, *J. Chem. Inf. Comput. Sci.* 41, 1457-1462, 2001.
32. D Bonchev, ON Temkin and D Kamenski, On the Complexity of Linear Reaction Mechanisms, *React. Kinet. Catal. Lett*. 15, 119-124, 1980.
33. D Bonchev, D Kamensky and ON Temkin, Complexity Index for the Linear Mechanisms of Chemical Reactions, *J. Math. Chem.* 1, 345-388, 1987.
34. K Gordeeva, D Bonchev, D Kamenski, and ON Temkin, Enumeration, Coding, and Complexity of Linear Reaction Mechanisms, *J. Chem. Inf. Comput. Sci*. 34, 244-247, 1994.
35. ON Temkin, AV Zeigarnik, and D Bonchev, *Chemical Reaction Networks. A Graph Theoretical Approach.* CRC Press, Boca Raton, FL, 1996.
36. F Harary, Graph Theory, 2nd printing, Addison-Wesley, Reading, MA, 1969.
37. F Harary, RZ Norman and D Cartwright, *Structural Models*: *An Introduction to the Theory of Directed Graphs,* Wiley, New York, 1965.
38. N Trinajstić, Chemical Graph Theory, 2nd ed., CRC Press, Boca Raton, FL, 1992.
39. HL Morgan, The Generation of a Unique Machine Description for Chemical Structure – A Technique Developed at Chemical Abstracts Service, *J. Chem. Docum.* 5, 107-113, 1965.

40. D Bonchev, AT Balaban and O Mekenyan, Generalization of the Graph Center Concept, and Derived Topological Indexes, *J. Chem. Inf. Comput. Sci*. 20, 106-113, 1980.
41. D Bonchev, The Concept for the Center of a Chemical Structure and Its Applications, *Theochem* 185, 155-168, 1989.
42. D Bonchev, O Mekenyan and AT Balaban, An Iterative Procedure for the Generalized Graph Center in Polycyclic Graphs, *J. Chem. Inf. Comput. Sci*. 29, 91-97, 1989.
43. MEJ Neuman, SH Strogatz and DJ Watts, Random Graphs With Arbitrary Degree Distribution and Their Applications, Santa Fe Institute, 2000, Working Paper 00-07-042.
44. M Gell-Mann, *The Quark and the Jaguar*, Freeman, New York, 1994, p.31.
45. D Bonchev, On the Complexity of Directed Biological Networks, *SAR QSAR Envir. Sci.* 14, 199-214, 2003.
46. D Bonchev. Complexity of Protein-Protein Interaction Networks, Complexes and Pathways, in *Handbook of Proteomics Methods*, M. Conn, ed. Humana, New York, 2003, p. 451-462.
47. D Bonchev, Complexity Analysis of Yeast Proteome Network, Chem. & Biodiversity, 1, 312-332, 2004.
48. Mathematical Chemistry Series, Vol. 7, *Complexity in Chemistry*, D Bonchev and DH Rouvray, Eds, Taylor & Francis, London, 2003.
49. S Nicolić, IM Tolić, N Trinajstić, and I Baučić, On the Zagreb Indices as Complexity Indices, *Croat. Chem. Acta* 73, 909-921, 2000.
50. M Randić and D Plavšić, On the Concept of Molecular Complexity, *Croat. Chem. Acta* 75, 107-116, 2002.
51. JR Platt, Prediction of Isomeric Differences in Paraffin Properties, *J. Phys. Chem.* 56, 328-336, 1952.
52. D Bonchev, On the Complexity of Platonic Solids, *Croat. Chem. Acta* 77, 167-173, 2004.
53. M Gordon and GR Scantleburry, Non-random Polycondensation: Statistical Theory of the Substitution Effect, *Trans. Faraday Soc.* 60, 604-621, 1964.
54. SH Bertz and WC Herndon, The Similarity of Graphs and Molecules. In: TH Pierce, and BA Hohne, Eds, *Artificial Intelligence Applications to Chemistry*, ACS, Washington, DC, 1986, p.169-175.
55. G Rücker and C Rücker, Automatic Enumeration of Molecular Substructures, *MATCH – Commun. Math. Comput. Chem.* 41, 145-149, 2000.
56. D Bonchev and N Trinajstić, Overall Molecular Descriptors. 3. Overall Zagreb Indices, *SAR QSAR Environ. Res.* 12, 213-235, 2001.
57. D Bonchev, Overall Connectivity –A Next Generation Molecular Connectivity, *J. Mol. Graphics Model.* 20, 55-65, 2001.
58. H Wiener, Structural Determination of Paraffin Boiling Points, *J. Am. Chem. Soc.* 69, 17-20, 1947.
59. H Wiener, Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure, *J. Phys. Chem.* 52, 1082-1089, 1948.
60. D Bonchev, The Overall Wiener Index - A New Tool for Characterization of Molecular Topology, *J. Chem. Inf. Comput. Sci*. 41, 582-592, 2001.

61. I Gutman, B Ruščić, N Trinajstić and CW Wilcox, Jr, Graph Theory and Molecular Orbitals. 12. Acyclic Polyenes, *J. Chem. Phys.* 62, 3399-3405, 1975.
62. S Nicolić, G Kovacevic, A Milicevic and N Trinajstić, The Zagreb Indices 30 Years After, *Croat. Chem. Acta* 76, 113-124, 2003.
63. DJ Watts and SH Strogatz, Collective Dynamics of "Small-World" Networks, *Nature* 393, 440-442, 1998.
64. AL Barabási, *Linked. The New Science of Networks.* Perseus, Cambridge, MA, 2002.
65. AL Barabási and R Albert, Emergence of Scaling in Random Networks, *Science* 286, 509-512, 1999.
66. SN Dorogovtsev and JFF Mendes, Evolution of networks, *Adv. Phys.* 51, 1079-1187, 2002.
67. http://www.santafe.edu/sfi/publications/working-papers.html.
68. T Ito, T Chiba, R Ozava, M Yoshida, M Hattori and Y Sasaki, A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome, *Proc. Natl. Acad. Sci. USA* 98, 4569-4574, 2001.
69. G Weng, US Bhala and RYyengar, Complexity in Biological Signaling Systems, *Science* 284, 92-96, 1999.
70. A Wagner, The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes, *Mol. Biol. Evol.* 18, 1283–1292, 2001.
71. L Giot *et al*, A Protein Interaction Map of Drosophila Melanogaster, *Science* 302, 1727-1736, 2003.
72. AC Gavin *et al*, Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes, *Nature* 415,141-147, 2002.
73. TI Lee et al, Transcriptional Regulatory Networks in Saccharomyces Cerevisiae, *Science*, 298, 799-804, 2002.
74. N Friedman, Inferring Cellular Networks Using Probabilistic Graphical Models, *Science*, 303, 799-805, 2004.
75. AHY Tong *et al,* Global Mapping of the Yeast Genetic Interaction Network, *Science*, 303, 606-813, 2004.
76. H Jeong, B Tombor, Z Albert and AL Barabási, The Large-Scale Organization of Metabolic Networks, *Nature* 407, 651-654, 2000.
77. A Wagner and DA Fell, The Small World Inside Large Metabolic Networks, *Proc. Roy. Soc. London B* 268, 1803-1810, 2001
78. H Ma and AP Zeng1, Reconstruction of Metabolic Networks from Genome Data and Analysis of Their Global Structure for Various Organisms, *Bioinformatics* 19, 270–277, 2003.
79. C Koch and G Laurent, Complexity and the Nervous System, *Science* 284, 96-98, 1999.
80. S Karabunarliev and D Bonchev, Grafman software package, unpublished.
81. SN Dorogovtzev, JFF Mendes and AN Samukhin, Giant Strongly Connected Component of Directed Graphs, arXiv I cond-mat/0103629 v1 Mar 2001.
82. SH Yook, H Jeong and AL Barabási, Modeling the Internet's Large-Scale Structure, *Proc. Natl. Acad. Sci.* 99, 13382-13386, 2002.

83. JA Dunne, RJ Williams and ND Martinez, Networks Topology and Biodiversity Loss in Food Webs: Robustness Increases With Connectance, Santa Fe Institute Working Paper 02-03-013, 2002.
84. R Albert, H Jeong and AL Barabási, Error and Attack Tolerance of Complex Networks, *Nature* 406, 378-382, 2000.
85. S Maslov and K Sneppen, Specificity and Stability in Topology of Protein Networks, *Science* 296, 309-313, 2002.

Table 1. The newly defined complexity index B matches well the complexity ordering of six-vertex graphs with the same connectedness as produced by four other complexity measures

Graph	**A/D**	**B = $\Sigma a_i/d_i$**	**SC**	**OC**	**TWC**	**I_{vd}**
17	0.250	1.833	62	535	852	15.61
18	0.231	1.636	56	475	754	14.75
19	0.231	1.567	52	426	598	14.00
20	0.231	1.464	43	329	450	12.75
21	0.222	1.558	53	444	708	13.75*
22	0.222	1.544	49	394*	662	14.00*
23	0.222	1.483	49	396*	556	13.51
24	0.222	1.464	37	264	372	12.00
25	0.214	1.439	44*	343*	564	13.51
26	0.214	1.417	48*	386*	540	13.51
27	0.207	1.408	45	354	602	13.51
28	0.207	1.354	42	318	480	12.75
29	0.194	1.260	37	266	490	12.75

Table 2. Adjacency, Average Vertex Degree, and Connectedness of the Nine Functional Groups of Protein Complexes in Saccharomyces Cerevisiae (calculated from data of Gavin *et el.*[72])

Protein Functional Group	V	V^a	A^b	$\langle a_i \rangle$	$Conn$	WA^c	$\langle w_i \rangle$	$Wconn$
RNA Metabolism	28	25	212	7.57	0.280	1124	40.14	1.487
Transcription/DNA Maintenance/Chromatin	55	44	468	8.50	0.158	1076	19.56	0.362
Protein Synthesis and Turnover	33	21	92	2.79	0.087	250	7.58	0.237
Membrane Biogenesis	20	11	40	2.00	0.106	44	2.20	0.116
Intermediate & Energy Metabolism	43	21	86	2.14	0.051	104	2.42	0.058
Protein RNA/Transport	12	6	12	1.00	0.091	20	1.67	0.152
Signaling	20	-	14	0.70	0.037	-	-	-
Cell Cycle	12	-	6	0.50	0.045	-	-	-
Cell Polarity & Structure	8	-	2	0.25	0.036	4	0.50	0.071

[a]The number of vertices in the giant component. No such component is available in the last three groups. [b]The connectivity measures are calculated for the entire network, not for the giant component only. [c]The calculations of the weighted indices is done with eqs. (9) and (10), while those of the non-weighted indices by eqs (2) and (4).

Table 3. Complexity Measures of Six Functional Groups of Protein Complexes in Saccharomyces Cerevisiae (calculated from data of Gavin *et el.*[72]): Second- and Third-Order Subgraph Count, First- and Second-Order Overall Connectivity, Information on the Vertex Degree Distribution, and A/D Complexity Index

Protein Functional Group[a]	^{2}SC	^{3}SC	^{1}OC	^{2}OC	$I_{vd,n}$	A/D
RNA Metabolism	7.396	27.472	7.868	33.843	0.627	1.083
Transciption/DNA Maintenance/Chromatin	0.605	0.650	0.631	0.729	0.522	0.289
Protein Synthesis and Turnover	0.675	0.591	0.844	1.100	0.546	0.268
Membrane Biogenesis	0.200	0.095	0.224	0.115	0.422	0.216
Intermediate & Energy Metabolism	0.107	0.043	0.117	0.055	0.421	0.112
Protein RNA/Transport	0.517	0.312	0.640	0.448	0.477	0.385

[a]The functional groups of protein complexes involved in signaling, cell cycle, and cell polarity & structure are omitted, because they lack a giant component. The calculations are performed by the Grafman software,[80] making also use of eqs. (21, 23, 31, 32).

FIGURE CAPTION

Figure 1. a) A disconnected graph with three components. b) A simple connected undirected graph. c) A directed graph. d) A complete graph with three cycles (the enveloping cycle is not counted, because it is not an independent cycle). e) A multigraph with a loop: 1, edge; 2, double edge; 3, loop. f) A weighted graph.

Figure 2. The undirected graph **1**, the directed graph **2**, their adjacency matrices $\boldsymbol{A}(\mathbf{1})$ and $\boldsymbol{A}(\mathbf{2})$, and total adjacencies $A(\mathbf{1})$ and $A(\mathbf{2})$, respectively.

Figure 3. Iterative calculation of the first- and second-order extended connectivity of graph **2** (The null-order is identical to the total adjacency of the graph).

Figure 4. Distance matrices $\boldsymbol{D}(\mathbf{1})$ and $\boldsymbol{D}(\mathbf{2})$, total distances $D(\mathbf{1})$ and $D(\mathbf{2})$, average distance degrees $\langle d_i \rangle$, and average distances $\langle d \rangle$, of the undirected graph **1**, the directed graph **2**, respectively.

Figure 5. Which graph is more complex: the totally disconnected graph **a** or the complete graph **b**?

Figure 6. The binomial distribution of vertex degrees in random graphs is used as an argument that complexity of graphs passes through a maximum with the increase in connectivity.

Figure 7. Thirteen graphs with five vertices ordered according to their increasing complexity, adequately matched by the values of the information index for the vertex degree distribution.

Figure 8. The larger complexity of graph **1** as compared to graph **16** is demonstrated by the total, average and normalized number of two-edge subgraphs ${}^{2}SC$, ${}^{2}SC_a$, and ${}^{2}SC_n$, respectively, as well as by the graph connectedness *Conn*, which is identical to the normalized number of edges, ${}^{1}SC_n$.

Figure 9. Thirteen graphs with five vertices ordered according to their increasing complexity, adequately matched by the values of the subgraph count *SC*, overall connectivity *OC*, and the total walk count *TWC*.

Figure 10. With few exceptions for the *A/D* and I_{vd} indices all the six complexity measures match the increase in complexity of graphs **3** through **15.**

Figure 11. Thirteen graphs with six vertices and sis edges used as a test for the sensitivity of the complexity measures

Figure 12. Special subclasses of directed graphs belonging to the classes of monocyclic, complete, path, and star graphs, respectively. The DC_V and DK_V subclasses shown have a complete vertex accessibility. Directed star graphs DS_V have the highest accessibility when a half of the arcs are incoming to and the other half of the arcs are outgoing from the central vertex.

Figure 13. The network of the protein complexes functional group of RNA metabolism in Saccharomyces Cerevisiae. The complexes sequential numbers and connectivity table are those from Gavin *et al.*[72] A pair of vertices are connected by an edge when they share at least one protein. (Not shown are three complexes that do not share any proteins). The high complexity of the network indicates the high stability of the RNA metabolism against random attacks and mutations.

Figure 14. A typical structure of a complex directed graph

Figure 15. The connectivity of the StMartin island food web[83] is illustrated in two directed graphs formed by the hierarchically ordered layers A to F. The trophic species of

each layer (numbered after ref. 83) can eat only downstream species and, in few cases, species of their own layer. The connectivity shown explicitly in the upper (unweighted) graph is that between the pairs of neighboring layers only. The edges of the lower (weighted) graph show the total number of interactions between all pairs of layers. The calculations of the complexity measures of the St. Martin's food web, however, are made proceeding from the entire directed graph with its 42 species and 205 directed interactions.

Figure 16 Complexity comparison of seven food webs (data from Dunne, Williams, and Martinez[83]) show the Skipwithpond and the Coachella Valley food webs to be the most complex ones, and the Canton Creek and Ythan Estuary to be the least complex ones. Complexity measures 1 to 6 correspond to connectedness (connectance), second- and third-order subgraph count, and first-, second-, and third-order overall connectivity.[24-29]

Figure 17. Stability analysis of the Ythan Estuary Food Web. The web splits into two pieces after eliminating the 13 highest connected vertices. The complexity measures used are the connectedness, the second-order subgraph count, and the first order overall connectivity.

FIGURES

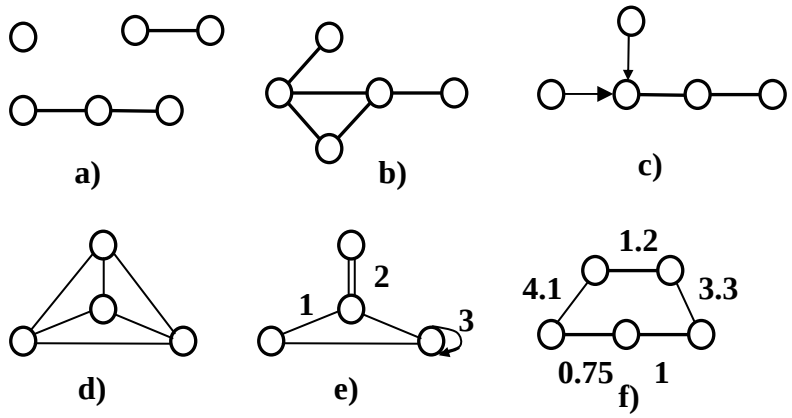
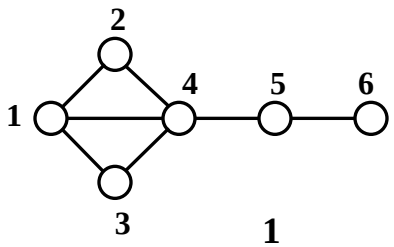
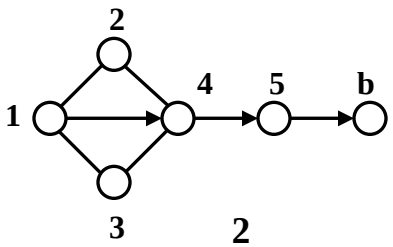


Fig. 1



1



2

$\boldsymbol{A}(\mathbf{1}) =$

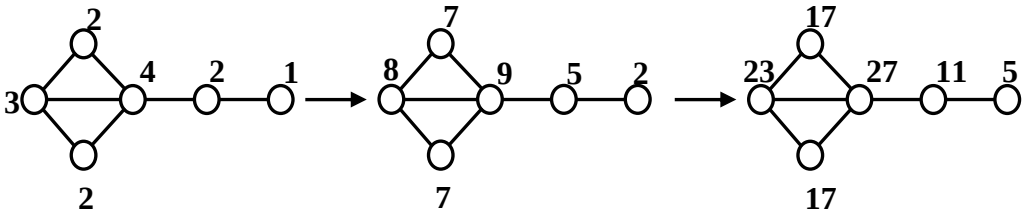
v	1	2	3	4	5	6	a_i
1	0	1	1	1	0	0	3
2	1	0	0	1	0	0	2
3	1	0	0	1	0	0	2
4	1	1	1	0	1	0	4
5	0	0	0	1	0	1	2
6	0	0	0	0	1	1	1

$\boldsymbol{A}(\mathbf{2}) =$

v	1	2	3	4	5	6	a_i
1	0	1	1	1	0	0	3
2	1	0	0	1	0	0	2
3	1	0	0	1	0	0	2
4	0	1	1	0	1	0	3
5	0	0	0	0	0	1	1
6	0	0	0	0	0	0	0

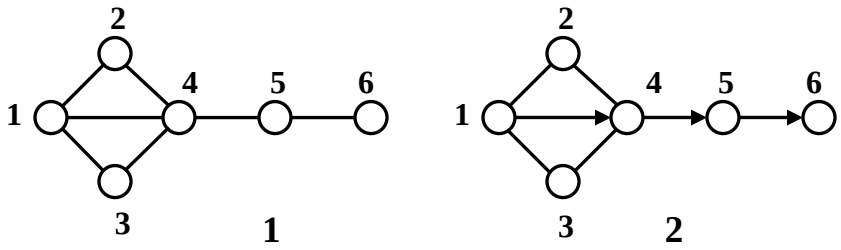
$A(\mathbf{1}) = 14$ $A(\mathbf{2}) = 11$

Fig. 2



${}^0EC = 14$ **${}^1EC = 38$** **${}^2EC = 100$**

Fig. 3.



***D*(1) =**

V	1	2	3	4	5	6	d_i
1	0	1	1	1	2	3	8
2	1	0	2	1	2	3	9
3	1	2	0	1	2	3	9
4	1	1	1	0	1	2	6
5	2	2	2	1	0	1	8
6	3	3	3	2	1	0	12

***D*(2) =**

V	1	2	3	4	5	6	d_i
1	0	1	1	1	2	3	8
2	1	0	2	1	2	3	9
3	1	2	0	1	2	3	9
4	2	1	1	0	1	2	7
5	-	-	-	-	-	0	1	1
6	-	-	-	-	-	-	0	0

$D(G) = 52$, $\langle di \rangle = 8.67$, $\langle d \rangle = 1.73$
$D(DG) = 34$, $\langle di \rangle = 5.67$, $\langle d \rangle = 1.62$

Fig. 4

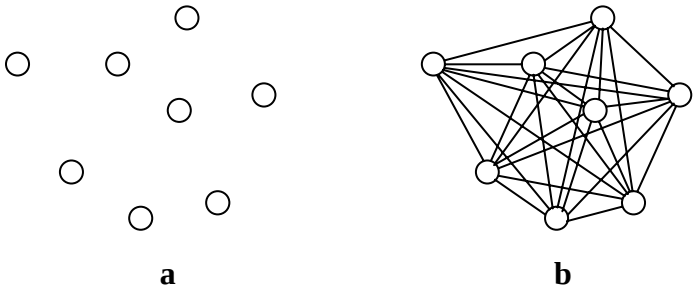


Fig. 5

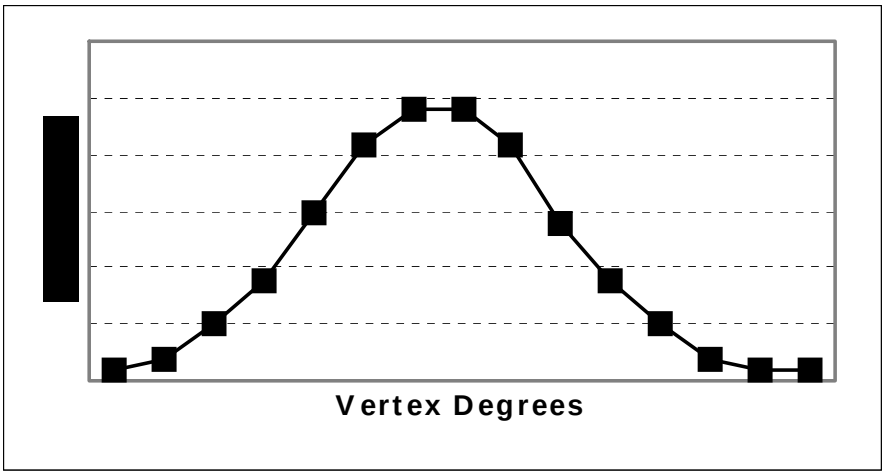


Fig. 6

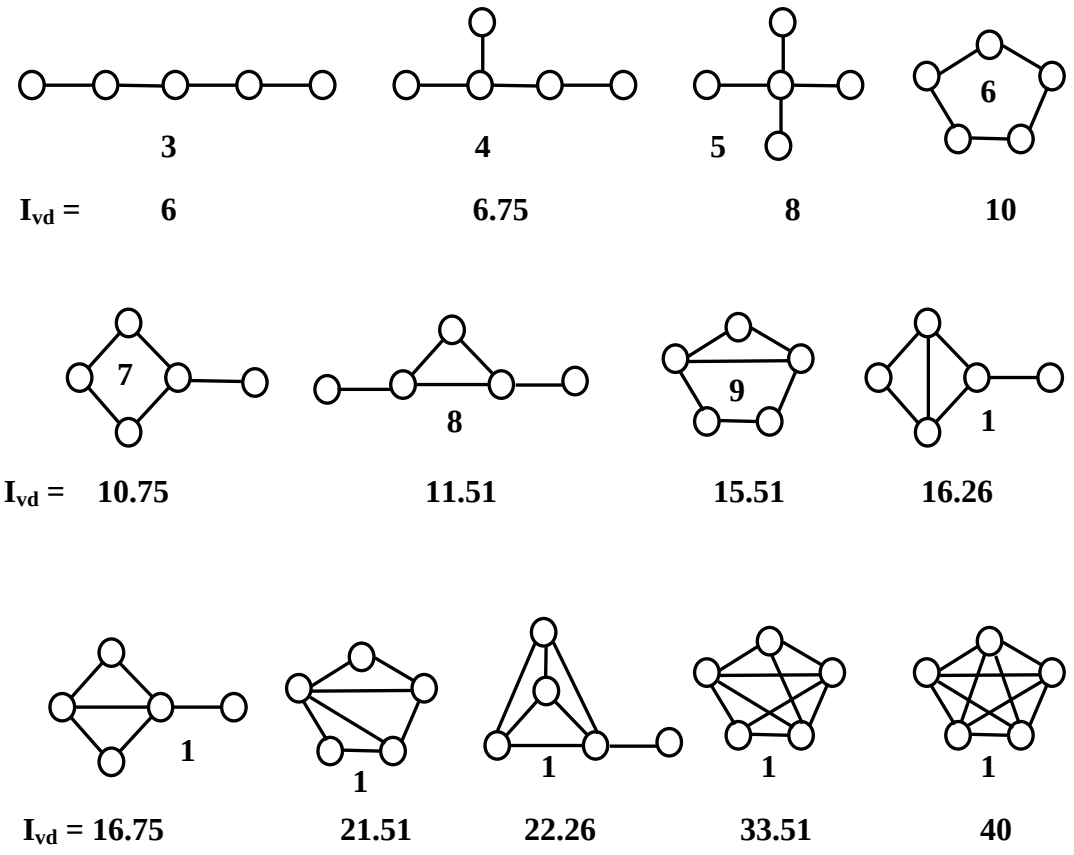
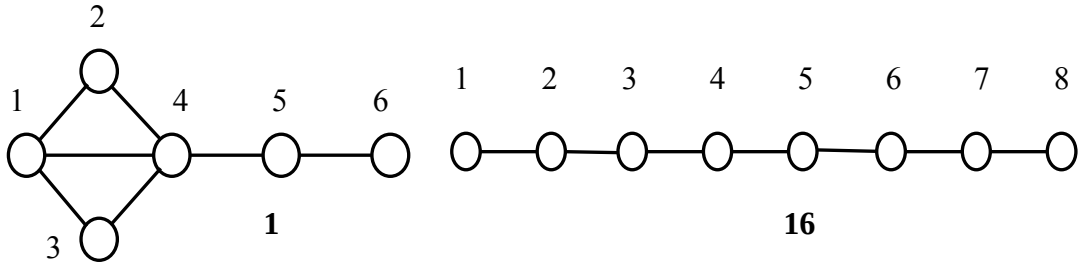

3
4
5
6
I_{vd} = 6
6.75
8
10
7
8
9
1
I_{vd} = 10.75
11.51
15.51
16.26
1
1
1
1
1
I_{vd} = 16.75
21.51
22.26
33.51
40

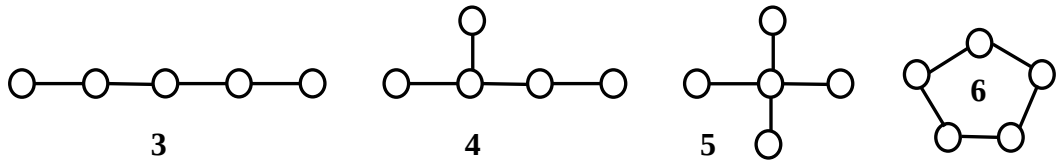
Figure 7



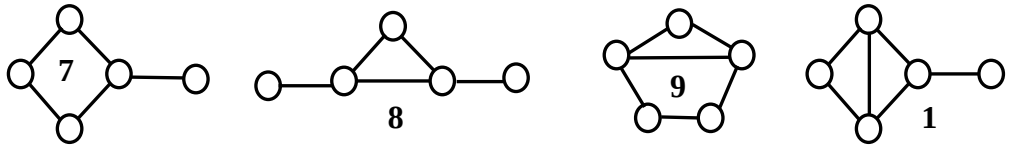
Graph **1:** 124, 134, 142, 143, 145, 213, 214, 243, 245, 314, 345, 456
***E* = 7, ^{2}SC** = 12, $^{2}SC_{a}$ = 2, $^{2}SC_{n}$ = 0.5**, *Conn* =** $^{1}SC_{n}$ = 0.233

Graph **16:** 123, 234, 345, 456, 567, 678
E = 7, ^{2}SC = 6, $^{2}SC_{a}$ = 0.75, $^{2}SC_{n}$ = 0.036, ***Conn*** = $^{1}SC_{n}$ = 0.125

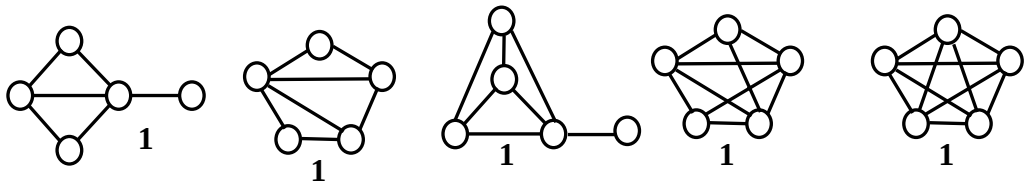
Figure 8



SC =	11	17	20	26
OC =	32	76	100	160
TWC =	58	106	140	150



SC =	29	31	54	57
OC =	190	212	482	522
TWC =	178	214	300	350



SC =	61	114	119	477	973
OC =	566	1316	1396	7806	18180
TWC =	337	538	608	1200	1700

Figure 9

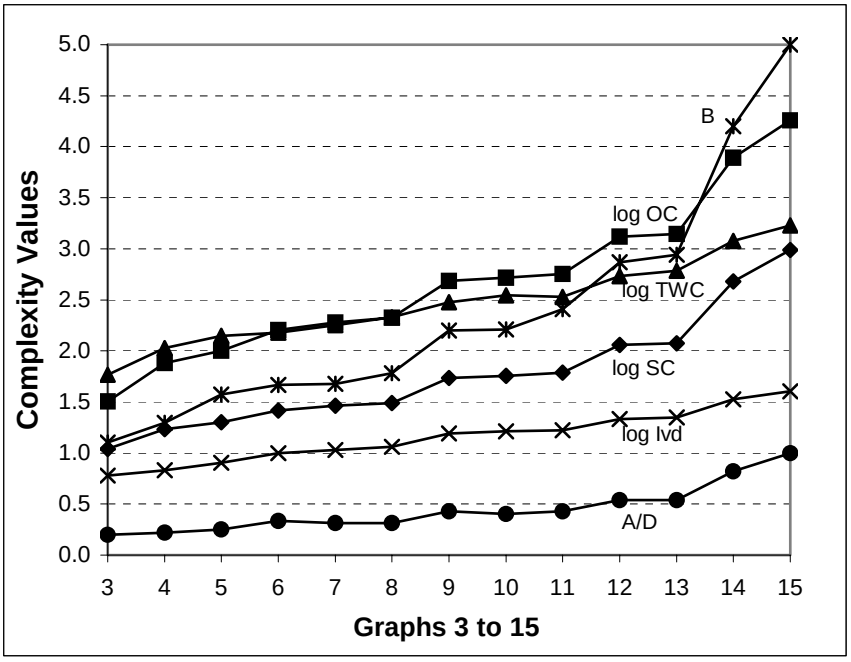

5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Complexity Values
3
4
5
6
7
8
9
10
11
12
13
14
15
Graphs 3 to 15
B
log OC
log TWC
log SC
log lvd
A/D

Figure 10

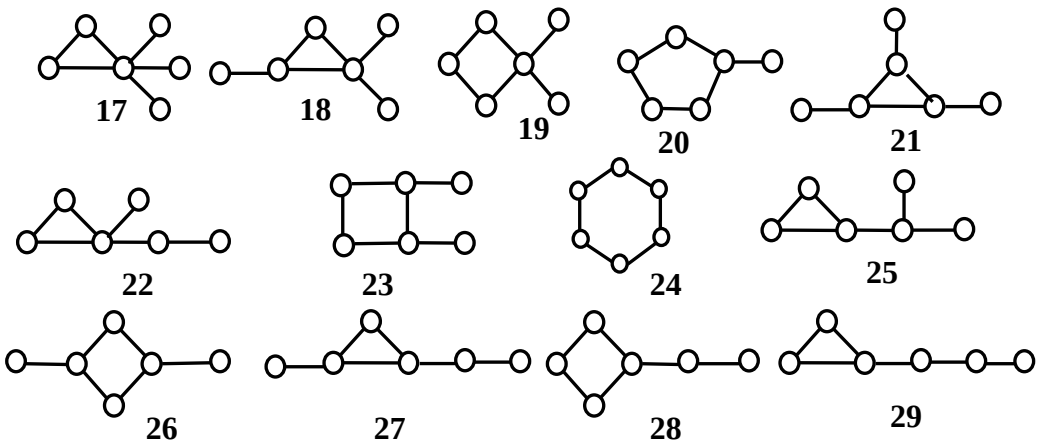


Figure 11

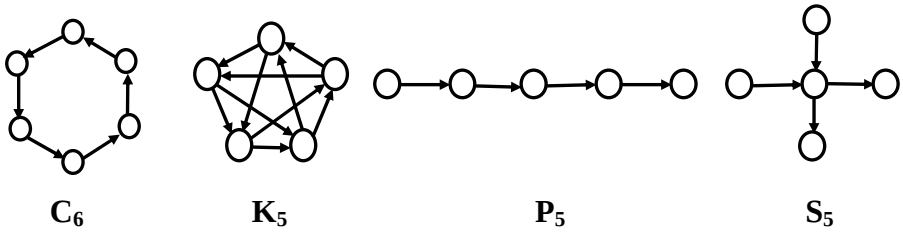


Figure 12

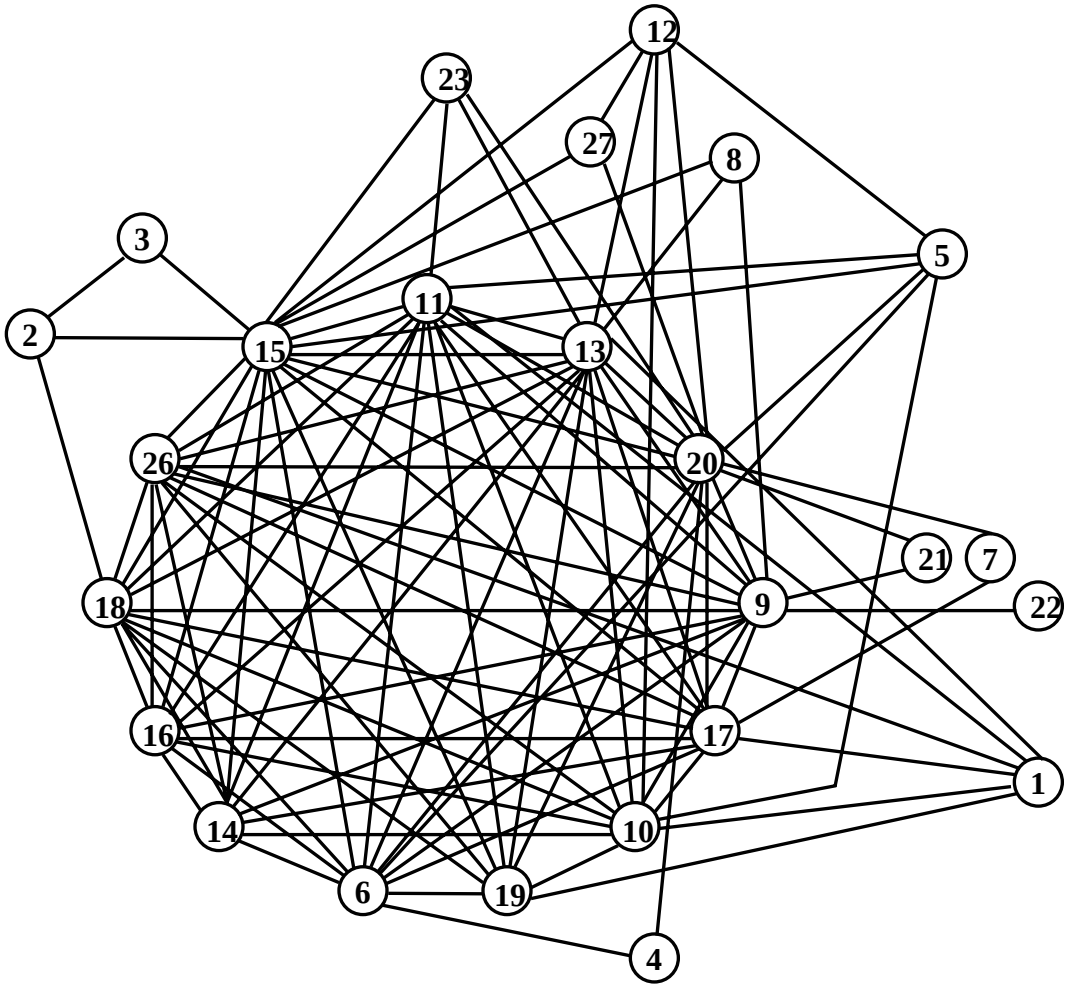


Figure 13

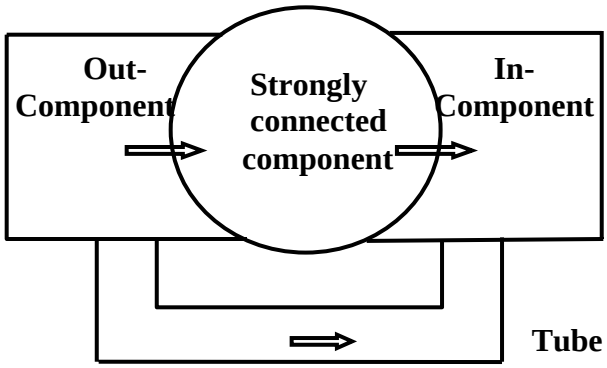
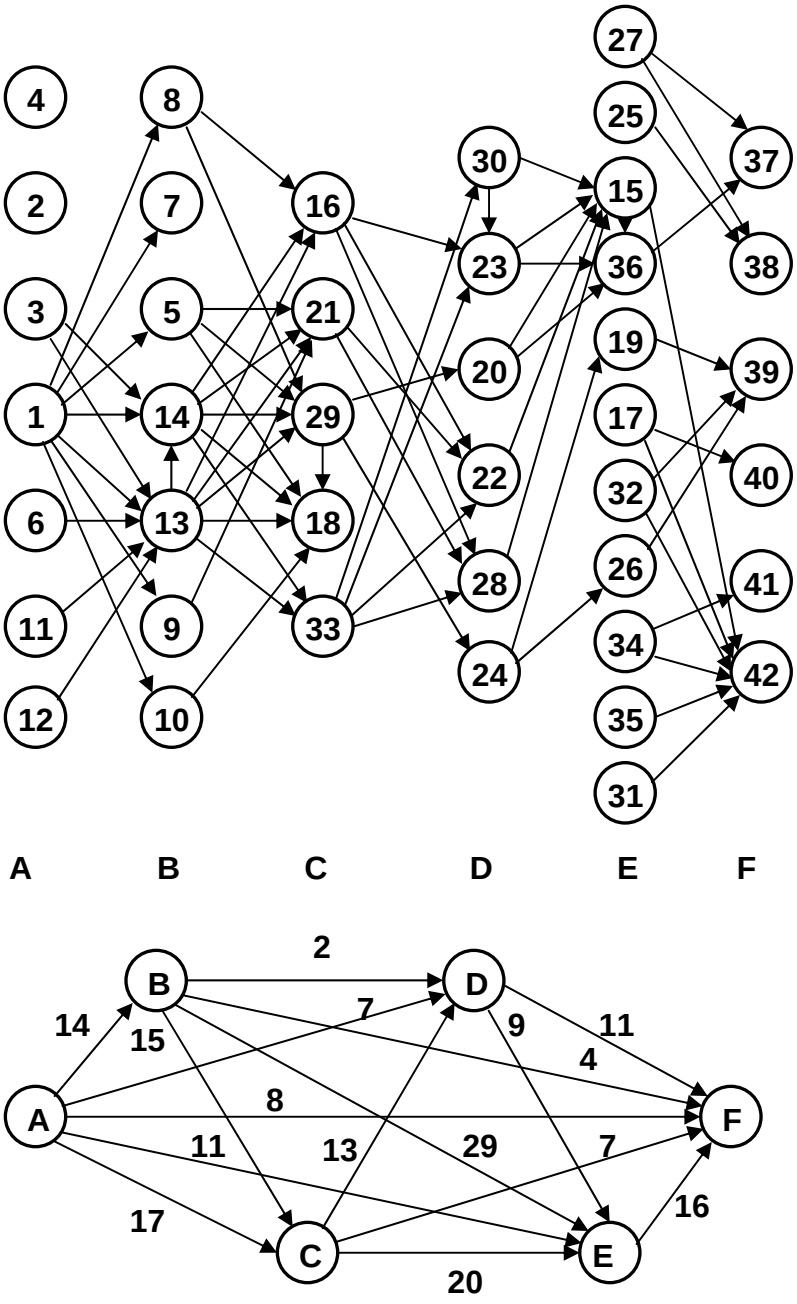

Out-
Component
Strongly
connected
component
In-
Component
Tube

Figure 14


4
2
3
1
6
11
12
8
7
5
14
13
9
10
16
21
29
18
33
30
23
20
22
28
24
27
25
15
36
19
17
32
26
34
35
31
37
38
39
40
41
42
A
B
C
D
E
F

A
B
C
D
E
F
2
14
15
7
8
11
13
29
9
11
4
7
17
16
20

Figure 15

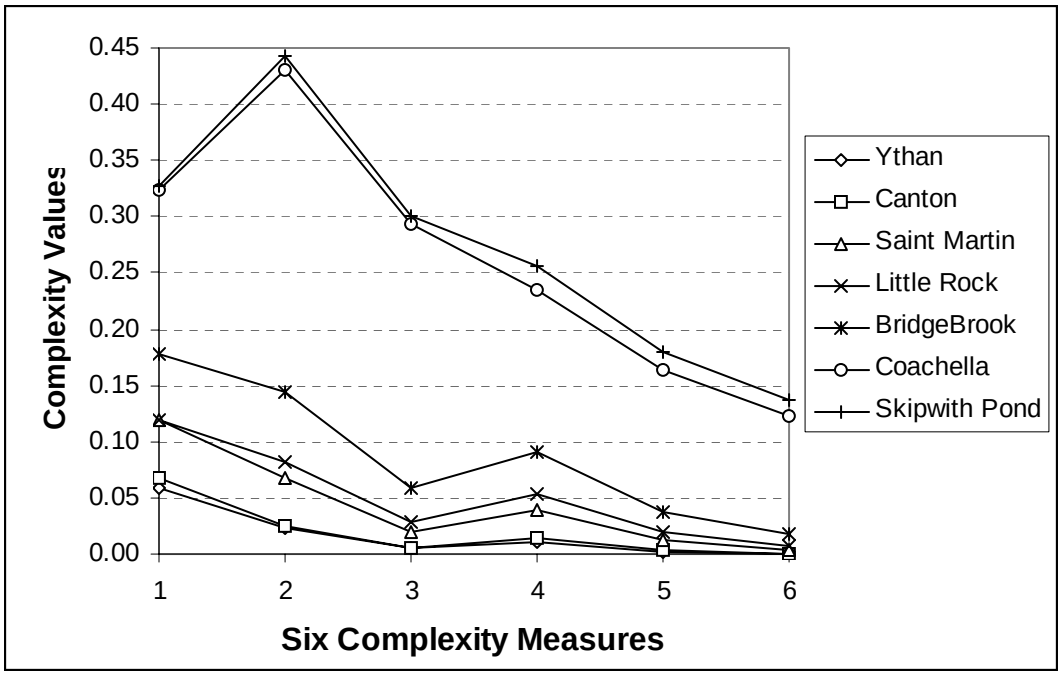

0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Complexity Values
1
2
3
4
5
6
Six Complexity Measures
Ythan
Canton
Saint Martin
Little Rock
BridgeBrook
Coachella
Skipwith Pond

Figure 16

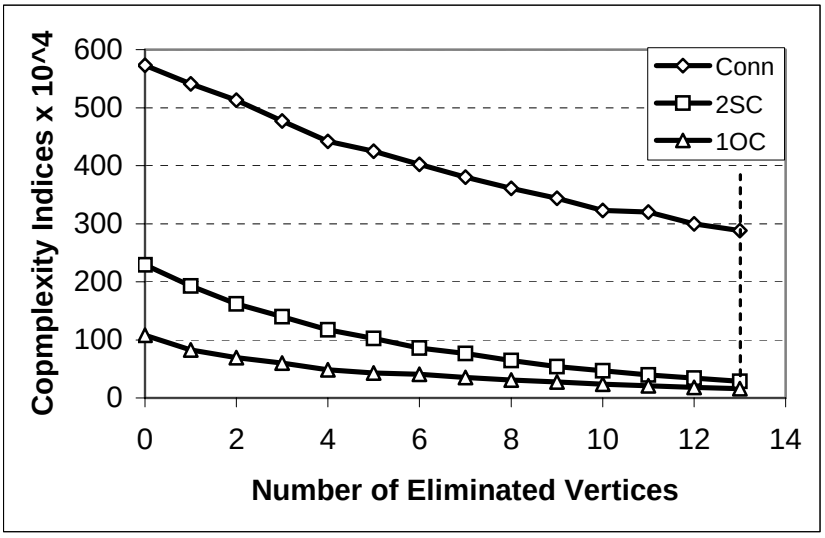
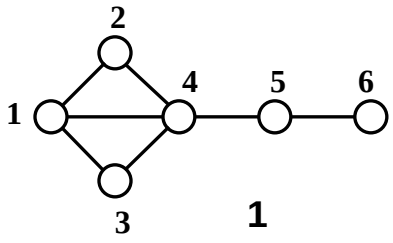


Figure 17



Scheme 1

To be inserted in subsection 5.3.6!!