Silesian J. Pure Appl. Math.
vol. 7, is. 1 (2017), 21–63

Piotr WILCZEK

Computer Laboratory, Poznań, Poland

NOVEL CENTRALITY MEASURES AND DISTANCE-RELATED TOPOLOGICAL INDICES IN NETWORK DATA MINING

Abstract. The present work proposes two new Euclidean distance functions, six new centrality measures as well as several new entropies definable on any complex network. It is demonstrated on four spatial and two social real-world datasets that these concepts are applicable in network data mining. Also, several new topological indices are introduced and their basic computational properties are established.

1. Introduction

Since the beginning of the twenty-first century, *network science*, the discipline whose main objective is to analyze network data, has become more and more popular in diverse fields of science and engineering. *Networks* (or more formally *graphs*) constitute a useful mathematical representation of a large variety of complex systems, from large online social networks (like Facebook) to chemical (or biochemical), ecological and infrastructural systems [17, 19, 33]. Roughly speaking, researchers in the field of *network data mining* try to develop mathematical models that discover patterns in interaction between different entities. They aim

2010 Mathematics Subject Classification: 05C12, 92E10.
Keywords: complex networks, centrality measures, topological indices, Chemical Graph Theory.
Corresponding author: P. Wilczek (piotr.wilczek.net@onet.pl).
Received: 30.06.2017.

at scrutinizing and extracting information from complex relational data in order to obtain a coherent description of physical/social reality or technological phenomena.

Countless empirical evidences gathered from real data suggest that a large amount of natural and human-made systems are organized under the form of complex networks with many distictive topological properties. These macroscopic features have been analyzed from the point of view of mathematics and statistical physics [1]. For instance, many complex networks exhibit the *scale-free property* and/or the *small world property*. Namely, A.-L. Barabási and R. Albert found that the degree distribution of numerous real-world networks is far from normal. In [3], they documented that the degree distribution of many natural networks follows a power law, that is $P(k) \sim ak^{-\gamma}$ where a is a constant and γ is positive exponent (it was determined empirically that γ varies between 2 and 3 for the majority of real networks). Here, $P(k)$ is referred to the fraction of vertices having degree equal to k. A complex network whose degree distribution follows a power law is called *scale-free*. Such a network is characterized by the fact that the owerwhelming majority of its nodes are connected to a relatively small number of other nodes with the exception of the so-called hub vertices, that possess an extremely high connectivity. In turn, a complex network with the small world property is characterized by the fact that its average path length ($\langle l \rangle$) is small compared to the size of this network [7, 9, 33, 44].

It should be emphasized here that such complex systems as transportation networks, mobile phone networks, river networks, power grids and water distribution networks are all instances of networks where space is relevant and topology alone does not provide all the information needed for a proper understanding of the nature of these objects. Consequently, it is possible to single out a separate class of complex networks, namely *spatial* networks [4]. These networks are characterized by the fact that their vertices occupy a precise position in two or three dimensional Euclidean space and their links constitute real physical connections. An analysis of the structure of such large spatial entities is important from a theoretical as well as practical point of view. Many studies from such disciplines as geography or urbanism are devoted to a deeper understanding of complex networks embedded in the real space. Also, note that progress in network data mining is parallel to the development of large virtual social networks. Roughly speaking, a social network is a collection of people or groups of people with some pattern of interrelatedness or links between them [33]. Such networks are scale-free and have the small world property.

It should also be highlighted that traditional data mining isuues (for instance, association rule mining or classification) try to extract some informative patterns based on individual data items. On the other hand, *network data mining* tries to discover structured relationships between different data objects, thereby describing emergent network patterns in complex relational datasets. This branch of data analysis uses some conceptual tools borrowed from statistical physics in order to extract knowledge from large numbers of individual datapoints [1].

The present paper introduces two new Euclidean distance functions definable on any complex network. It will be demonstrated on four spatial and two social networks that these novel notions are applicable in exploring real-world phenomena.

This article is organized as follows. Section 2 introduces two novel distance matrices describing complex networks. Section 3 introduces six new centrality metrics. Section 4 defines several novel centrality-based network complexity measures. In turn, some useful definitions of several new distance-related network invariants are included in Section 5. Sections 6 and 7 contain several methodological hints and many numerical results of experiments conducted on six exemplary and two randomly generated complex networks as well as on one dataset of all exhaustively generated small networks possessing up to 7 nodes. Section 7 also contains two examples of applications of the cencepts introduced in Sections 3, 4 and 5 in data mining. Section 8 contains some final remarks.

2. Two novel distance matrices associated with a complex network

In this paper, it is assumed that all considered complex networks are modelled by simple graphs of the general form $G = (V(G), E(G))$ where $V(G) = \{v_1, v_2, ..., v_n\}$ is the vertex set and $|E(G)| = m$ is the edge set. The *geodesic* (*topological*) *distance* between two vertices $v_i, v_j \in V(G)$, denoted by $d_G(v_i, v_j)$, is identified with the number of edges in any shortest path connecting them [25,43]. Note that it can be easily demonstrated that the shortest path distance does not satisfy Euclid's postulates [26]. For two nodes $v_i, v_j \in V(G)$, $v_i v_j$ means that v_i and v_j are *adjacent*, i.e., $v_i v_j \in E(G)$. The *neighborhood* of the vertex $v_i \in V(G)$, denoted by $N_G(v_i)$, is defined as $N_G(v_i) = \{w \in V(G) : wv_i \in E(G)\}$. The symbol k_i denotes the *degree* of the vertex v_i. Undoubtedly, $k_i = |N_G(v_i)|$. Given a complex network G with the vertex set $|V(G)| = n$, it is straightforward to

build its *adjacency matrix* $A(G)$ [25,43]. This network-theoretical object is identified with a real symmetric $n \times n$ two-dimensional array whose entries are given by the term $[A(G)]_{ij} = 1$ if $v_i v_j \in E(G)$ and $[A(G)]_{ij} = 0$ if $v_i v_j \notin E(G)$. In turn, the (*topological*, *geodesic*) *distance matrix* associated with a complex network $G = (V(G), E(G))$, denoted by $D(G)$, is a real symmetric $n \times n$ two-dimensional array whose elements $[D(G)]_{ij}$ are defined as $[D(G)]_{ij} = d_G(v_i, v_j)$ if $v_i \neq v_j$ and $[D(G)]_{ij} = 0$ if $v_i = v_j$ [25,43].

In 2010 M. Randić et al. [37] have introduced the novel distance matrix for networks. This two-dimensional array is referred to as the *natural distance matrix* and is denoted by $ND(G)$. Namely, for any complex network $G = (V(G), E(G))$ where $|V(G)| = n$, it is possible to interprete the rows of its adjacency matrix as points in the $n-$dimensional Euclidean space. Consequently, the natural distance between two nodes $v_i, v_j \in V(G)$, denoted by $d_G^N(v_i, v_j)$, is given by

$$d_G^N(v_i, v_j) := \left\{ \sum_{i=1}^{n} \left([A(G)]_{ik} - [A(G)]_{jk} \right)^2 \right\}^{\frac{1}{2}}.$$

The entries of $ND(G)$ matrix are identified with the natural distances between the points corresponding to nodes of G in the $n-$dimensional Euclidean space, i.e., $[ND(G)]_{ij} = d_G^N(v_i, v_j)$ if $v_i \neq v_j$ and $[ND(G)]_{ij} = 0$ if $v_i = v_j$. Thus, in this approach, the element $[ND(G)]_{ij}$ of the natural distance matrix is equal to the Euclidean distance between two adjacency (row) vectors of $A(G)$ in the $n-$dimensional space. Also in [37], it was demonstrated that the matrix entry $[ND(G)]_{ij}$ can be expressed as follows

$$[ND(G)]_{ij} := [k_i + k_j - |N_G(v_i) \cap N_G(v_j)|]^{\frac{1}{2}}.$$

Here, the term $|N_G(v_i) \cap N_G(v_j)|$ is equal to the number of nodes adjacent to both vertices v_i and v_j.

As mentioned previously, a complex network possessing n nodes can be represented by n points in the $n-$dimensional Euclidean space and can be described by the natural distance matrix $ND(G)$. In order to further generalize this geometric approach, we propose here to treat the rows of the adjacency matrix $A(G)$ associated with any complex network $G = (V(G), E(G))$ where $|V(G)| = n$ as points in some $n-$dimensional binary space $\{0,1\}^n$. In the present work, we will confine ourselves to two novel distances (and two novel distance structures) that can be defined on G. Thus, for any complex network G where $|V(G)| = n$, it is possible to single out the *Jaccard distance matrix*, denoted by $JD(D)$, whose

elements $[JD(G)]_{ij}$ are equal to the Jaccard distance between the nodes v_i and v_j if $v_i \neq v_j$ and to 0 otherwise. The *Jaccard distance* between two vertices v_i and v_j [21], denoted by $d_G^J(v_i, v_j)$, is defined as follows

$$d_G^J(v_i, v_j) := \left[1 - \frac{|N_G(v_i) \cap N_G(v_j)|}{k_i + k_j - |N_G(v_i) \cap N_G(v_j)|}\right]^{\frac{1}{2}}.$$

In this approach, the vertices $v_i, v_j \in V(G)$ are identified with their adjacency vectors and the distance between them is given by the Jaccard distance between these vectors in the binary space $\{0,1\}^n$. In turn, the *cosine* (or *Ochiai*) *distance matrix*, denoted by $\cos D(G)$, associated with any complex network $G = (V(G), E(G))$ where $|V(G)| = n$ has entries $[\cos D(G)]_{ij}$ equal to the cosine distance between two nodes v_i and v_j if if $v_i \neq v_j$ and to 0 otherwise. The *cosine* (or *Ochiai*) *distance* between two nodes v_i and v_j [21], denoted by $d_G^{\cos}(v_i, v_j)$, can be expressed by the subsequent formula

$$d_G^{\cos}(v_i, v_j) := \left[1 - \frac{|N_G(v_i) \cap N_G(v_j)|}{\sqrt{k_i k_j}}\right]^{\frac{1}{2}}.$$

Note that the term $\frac{|N_G(v_i) \cap N_G(v_j)|}{\sqrt{k_i k_j}}$ is equivalent to the cosine of the angle between two adjacency vectors corresponding to the nodes v_i and v_j. Also, in this case, two vertices $v_i, v_j \in V(G)$ are identified with their adjacency vectors and the distance between them is equal to the cosine distance between these vectors in the binary space $\{0,1\}^n$.

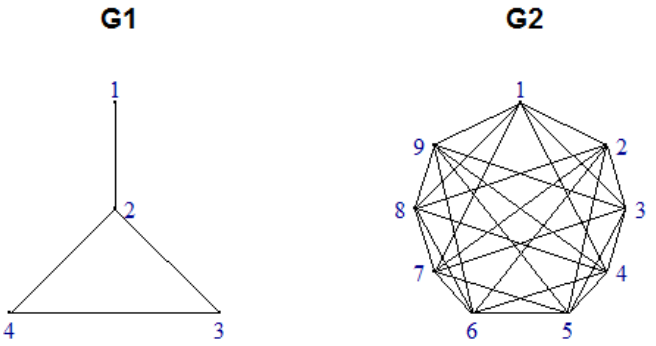

Fig. 1. Two sample networks G1 and G2

For instance, for the small network $G1$ in Figure 1, the topological, natural, Jaccard and cosine distance matrices have the following forms

$$D = \begin{bmatrix} & v_1 & v_2 & v_3 & v_4 \\ v_1 & 0 & 1 & 2 & 2 \\ v_2 & 1 & 0 & 1 & 1 \\ v_3 & 2 & 1 & 0 & 1 \\ v_4 & 2 & 1 & 1 & 0 \end{bmatrix}, \; ND = \begin{bmatrix} & v_1 & v_2 & v_3 & v_4 \\ v_1 & 0 & 2 & 1 & 1 \\ v_2 & 2 & 0 & 1.732 & 1.732 \\ v_3 & 1 & 1.732 & 0 & 1.414 \\ v_4 & 1 & 1.732 & 1.414 & 0 \end{bmatrix},$$

$$JD = \begin{bmatrix} & v_1 & v_2 & v_3 & v_4 \\ v_1 & 0 & 1 & 0.707 & 0.707 \\ v_2 & 1 & 0 & 0.866 & 0.866 \\ v_3 & 0.707 & 0.866 & 0 & 0.816 \\ v_4 & 0.707 & 0.866 & 0.816 & 0 \end{bmatrix},$$

$$\cos D = \begin{bmatrix} & v_1 & v_2 & v_3 & v_4 \\ v_1 & 0 & 1 & 0.541 & 0.541 \\ v_2 & 1 & 0 & 0.769 & 0.769 \\ v_3 & 0.541 & 0.769 & 0 & 0.707 \\ v_4 & 0.541 & 0.769 & 0.707 & 0 \end{bmatrix}.$$

Note that both new distance functions defined on any complex network satisfy the Euclidean axioms [21].

Based on these new distances and new distance matrices, it is possible to define several novel centrality measures as well as several novel distance-related topological indices.

3. New centrality measures

In this Section, after a short introduction, we will define six new vertex centrality measures. For any complex network $G = (V(G), E(G))$, a function $I_G : V(G) \to R$ is called a *vertex invariant* if for every $v \in V(G)$ the following condition is satisfied: $\forall G' \simeq G :\Longrightarrow I_{G'}(v') = I_G(v)$ where $v' \in V(G')$ and $v' = \varphi(v)$. Here, the relation $\simeq$ is an isomorphism between $V(G)$ and $V(G')$ such that if $v_i v_j \in E(G)$, then $\varphi(v_i)\varphi(v_j) \in E(G')$. However, if $I_G(v_i) = I_{G'}(v_i')$ for all $v_i \in V(G)$ and all $v_i' \in V(G')$, then the networks G and G' may or may not be isomorphic. In turn, if $I_G(v_i) \neq I_{G'}(v_i')$ for some $v_i \in V(G)$, then it can be

categorically concluded that G and G' are not isomorphic. For instance, almost all considered in the literature network centrality measures are vertex invariants [28].

Note that many questions that might be asked about a node in any complex network basically try to comprehend its "importance". A measure that indicates the relative importance of a node is known as the *centrality*. Such measures try to identify the most important vertices within a complex network. Because the term "importance" is ambiguous, a whole plethora of centrality measures has been proposed [28]. These statistical tools have found many applications in such fields as sociology, psychology, biology and computer science. Here, we will discuss two such concepts, starting with the *closeness centrality* (CC) as first introduces in sociology by A. Bavelas [6]. Note that Bavelas' closeness centrality is one of the most classical centrality measures in network data mining. It quantifies the centrality of a node in any complex network $G = (V(G), E(G))$ as the inverse of the sum of the distances to the other vertices in G. Its formula can be expressed as follows

$$CC(v_i) := \frac{1}{s(v_i)},$$

where $s(v_i) = \sum_{v_j \in V(G)} d_G(v_i, v_j)$. The quantity $s(v_i)$ is known as the *distance sum* of the vertex $v_i \in V(G)$ [43]. The closeness centrality is an indicator of the proximity between a vertex v_i and all other vertices in G. It quantitatively assesses the extent to which v_i is central to G. In this approach, the most important node (in terms of being close to most other vertices in G) is the node with the highest closeness centrality score. In turn, nodes with low closeness centrality scores are regarded as remote. However, this centrality measure has two main defects. First, it is well defined only for connected complex networks. Namely, any node that is unreachable from some other node has the closeness centrality score equal to zero. Second, even if the complex network is connected, the values of CC are dominated by distant vertices. To overcome these shortcomings, Y. Rochat proposed to use the harmonic mean of all shortest path distances [41]. This quantity is well defined even when some nodes are not connected. For any complex network $G = (V(G), E(G))$ where $|V(G)| = n$, the so-called *harmonic centrality* is expressed by the following formula

$$HC(v_i) := \frac{1}{n-1} \sum_{v_j \in V(G), v_j \neq v_i} \frac{1}{d_G(v_1, v_j)}.$$

In the literature, the normalization term $\frac{1}{n-1}$ is almost always omitted. In this study, we also omit this term. Therefore, the harmonic centrality has the form

$$HC\left(v_i\right) := \sum_{v_j \in V(G), v_j \neq v_i} \frac{1}{d_G\left(v_1, v_j\right)}.$$

This centrality was *implicitly* present in [24] as the "reciprocal distance sum".

Note that the closeness and harmonic centralities were introduced for the shortest path distance. In the following part of this Section, we will generalize these two notions to the natural, Jaccard and cosine distances. Thus, we will obtain three novel closeness-type and three novel harmonic-type centrality measures which in many cases exhibit better properties than their geodesic-based counterparts (cf. Section 7).

In order to present three new closeness-type centralities, let us define for a complex network $G = \left(V\left(G\right), E\left(G\right)\right)$ the following quantities

$$s^N\left(v_i\right) := \sum_{v_j \in V(G)} d_G^N\left(v_i, v_j\right) = \sum_{j=1}^{n} \left[ND\left(G\right)\right]_{ij},$$

$$s^J\left(v_i\right) := \sum_{v_j \in V(G)} d_G^J\left(v_i, v_j\right) = \sum_{j=1}^{n} \left[JD\left(G\right)\right]_{ij}$$

and

$$s^{\cos}\left(v_i\right) := \sum_{v_j \in V(G)} d_G^{\cos}\left(v_i, v_j\right) = \sum_{j=1}^{n} \left[\cos D\left(G\right)\right]_{ij}.$$

Here, $s^N\left(v_i\right)$, $s^J\left(v_i\right)$ and $s^{\cos}\left(v_i\right)$ denote the *natural*, *Jaccard* and *cosine distance sum* corresponding to the vertex $v_i \in V\left(G\right)$, respectively. Then, the subsequent expressions

$$NC\left(v_i\right) := \frac{1}{s^N\left(v_i\right)},$$

$$JC\left(v_i\right) := \frac{1}{s^J\left(v_i\right)}$$

and

$$\cos C\left(v_i\right) := \frac{1}{s^{\cos}\left(v_i\right)}$$

can be understood as the *natural*, *Jaccard* and *cosine closeness centrality* corresponding to the vertex $v_i \in V\left(G\right)$, respectively. Thus, based on the row sums of the natural, Jaccard and cosine distance matrices, we obtained three new closeness-type centrality measures.

On the other hand, based on the general definition of the harmonic centrality, it is possible to introduce for a complex network G three novel harmonic-type measures. They are expressed by the following formulae

$$NHC(v_i) := \sum_{v_j \in V(G), v_j \neq v_i} \frac{1}{d_G^N(v_i, v_j)},$$

$$JHC(v_i) := \sum_{v_j \in V(G), v_j \neq v_i} \frac{1}{d_G^J(v_i, v_j)}$$

and

$$\cos HC(v_i) := \sum_{v_j \in V(G), v_j \neq v_i} \frac{1}{d_G^{\cos}(v_i, v_j)}.$$

In this context, the quantities $NHC(v_i)$, $JHC(v_i)$ and $\cos HC(v_i)$ denote the *natural*, *Jaccard* and *cosine harmonic centrality* corresponding to the vertex $v_i \in V(G)$, respectively.

For instance, for the network $G1$ in Figure 1, the distance sums $s(v_i)$ and the Jaccard distance sums $s^J(v_i)$ have the following values: $s(v_1) = 5$, $s(v_2) = 3$, $s(v_3) = s(v_4) = 4$ and $s^J(v_1) \approx 2.414$, $s^J(v_2) \approx 2.732$, $s^J(v_3) = s^J(v_4) \approx 2.389$. Therefore, the values of CC and JC measures are as follows $CC(v_1) = \frac{1}{5}$, $CC(v_2) = \frac{1}{3}$, $CC(v_3) = CC(v_4) = \frac{1}{4}$ and $JC(v_1) \approx 0.4143$, $JC(v_2) \approx 0.366$, $JC(v_3) = JC(v_4) \approx 0.4186$. The values of the harmonic centrality for the network $G1$ are as follows $HC(v_1) = 2$, $HC(v_2) = 3$, $HC(v_3) = HC(v_4) = 2.5$ whereas the values of JHC measure for this network are given by $JHC(v_1) \approx 3.829$, $JHC(v_2) \approx 3.309$ and $JHC(v_3) = JHC(v_4) \approx 3.795$.

In turn, the nodes of the $G2$ network in Figure 1 can not be distinguished by the degree (DC), closeness (CC), harmonic (HC), betweenness (BC), eigenvector (EC) and PageRank (PRC) centrality measures. Namely, $DC(v_i) = 6$, $CC(v_i) = 0.1$, $HC(v_1) = 7$, $BC(v_i) = 1$, $EC(v_i) = 1$, $PRC(v_i) = 0.1111$ (with the damping factor equal to 0.85) for all $v_i \in V(G2)$. On the other hand, all six newly introduced centralities divide the vertex set of the network $G2$ into two equivalence classes $V_1 = \{v_1, v_3, v_5, v_6, v_8\}$ and $V_2 = \{v_2, v_4, v_7, v_9\}$. The values of these centrality measures for the partitions V_1 and V_2 are contained in Table 1. More numerical results confirming the higher specificity of these six newly introduced centrality measures are presented in Section 7.

Table 1
The values of six newly introduced centrality measures defined on the G2 network from Figure 1

Partition	NC	JC	$\cos C$	NHC	JHC	$\cos HC$
V_1	0.0601	0.1739	0.2083	4.0472	11.4691	14.0199
V_2	0.0606	0.1759	0.2101	4.1626	11.7362	14.4195

4. Some complexity measures associated with centralities

The problem to quantitatively assess the complexity of a network appears in varius scientific fields. This topic first appeared when investigating the complexity of biological and chemical networks. When analyzing the notion of complexity, *Information Theory* has been occupying the most noticeable position [14,43]. In this approach, an important question is to quantitatively evaluate the so-called *structural information content* of complex networks by applying *Shannon's information measure*. Besides describing biological or chemical systems, this paradigm of research was successfully applied in computer science, ecology, sociology, mathematical psychology, linguistics and physics (cf. [14] and the literature cited therein). Historically speaking, N. Rashevsky, R.H. McArthur, E. Trucco were the first who used Shannon's formula in order to define an *entropy* of a complex network (cf. [14, 43]). At present, there are many known complexity metrics based on Shannon's information measure. For instance, the *topological information content* developed by N. Rashevsky, the *symmetry index* for networks developed by A. Mowshowitz, the *chromatic information content* also developed by A. Mowshowitz, the *magnitude-based information indices* developed by D. Bonchev, the *vertex degree equality-based information index* also developed by D. Bonchev and the *overall information indices* also due to A. Bonchev (for all these measure cf. [14,43]). In this Section, we are going to introduce several information-theoretic network complexity measures based on the newly proposed centralities.

Note that any centrality measure C defined on a complex network $G = \big(V(G), E(G)\big)$ induces some equivalence relation $\simeq$ on the vertex set $V(G)$ given by the condition $v_i \simeq v_j \longleftrightarrow C(v_i) = C(v_j)$. Therefore, it is possible to obtain a partitioning of the set $V(G)$ where the resulting partitions are symbolized by $V_1, V_2, ..., V_k$. In this context, the entities $p_i = \frac{|V_i|}{|V|}$ where $1 \leq i \leq k$ constitute probabilities for each obtained partition V_i. Namely, it is apparent that $0 \leq p_i \leq 1$

and $\sum_{i=1}^{k} p_i = 1$. Consequently, the vector $P(G) = (p_1, p_2, ..., p_k)$ can be understood as a *finite probability distribution* of G and its Shannon's entropy I is given by

$$I(p) := -\sum_{i=1}^{k} p_i \log_2 p_i.$$

Thus, for any centrality measure C that induces a partition of the vertex set $V(G)$ into k many disjoint subsets of cardinality $|V_i|$, we obtain the following information-theoretic complexity measure

$$\overline{I}_C(G) := -\sum_{i=1}^{k} \frac{|V_i|}{|V|} \log_2 \left(\frac{|V_i|}{|V|} \right).$$

Here, the quantity $\overline{I}_C$ is termed the *mean information content* of the complex network G with respect to the centrality measure C (for these measures cf. [14,43] and the literature cited therein). Consequently, we have arrived at the definitions of eight complexity measures: $\overline{I}_{CC}$, $\overline{I}_{NC}$, $\overline{I}_{JC}$, $\overline{I}_{\cos C}$, $\overline{I}_{HC}$, $\overline{I}_{NHC}$, $\overline{I}_{JHC}$ and $\overline{I}_{\cos HC}$. Such constructed invariants are known as *partition-dependent complexity measures* of G.

In turn, the so-called *partition-independent complexity measures* of any complex network $G = (V(G), E(G))$ (where $|V(G)| = n$) with respect to some centrality meaure C are defined as follows. Let C be an aribtrary centrality measure imposed on $V(G)$. Then, it is possible to define for every vertex $v_i \in V(G)$ the following quantity

$$p^C(v_i) := \frac{C(v_i)}{\sum_{j=1}^{n} C(v_j)}.$$

Because the subsequent equation $p^C(v_1) + p^C(v_2) + ... + p^C(v_n) = 1$ is *a priori* valid, the quantities $p^C(v_i)$ can be understood as vertex probabilities. Consequently, the vector $P(G) = \left(p^C(v_1), p^C(v_2), ..., p^C(v_n)\right)$ constitutes a *finite probability distribution*. Its entropy is given by the following expression

$$I_C(G) := -\sum_{i=1}^{n} \frac{C(v_i)}{\sum_{j=1}^{n} C(v_j)} \log_2 \left(\frac{C(v_i)}{\sum_{j=1}^{n} C(v_j)} \right).$$

Such defined complexity measures representing the so-called *structural information content* of G are known as *parametric* network entropies [14, 43]. Note that in this approach, instead of inducing partitions of the vertex set $V(G)$ using some equivalence relation, informtaion-theoretic complexity measures for complex networks are defined by assigning a probability value to each node of a network. In our context, such probabilities values are given by using four closeness-type and four harmonic-type centralities. Hence, we obtained eight network-level invariants: I_{CC}, I_{NC}, I_{JC}, $I_{\cos C}$, I_{HC}, I_{NHC}, I_{JHC} and $I_{\cos HC}$.

For instance, for the network $G1$ in Figure 1, the closeness and Jaccard closeness centralities divide the vertex set $V(G1)$ into 3 equivalence classes: $V_1 = \{v_1\}$, $V_2 = \{v_2\}$ and $V_3 = \{v_3, v_4\}$. Therefore,

$$\overline{I}_{CC}(G1) = \overline{I}_{JC}(G1) \approx$$
$$\approx -\left(\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \left(\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \left(\frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right) \approx 1.5.$$

Several numerical results concerning eight partition-dependent as well as eight partition-independent complexity measures evaluated on six real-world and two randomly generated networks are included in Section 7. Also, their discriminatory abilities are studied in Section 7.

5. New topological indices

Let $\mathcal{G}$ be the class of all networks. A function $TI : \mathcal{G} \to \mathbb{R}$ is termed a *topological index* (or *network invariant*) if and only if the following condition is satisfied: $\forall G' \simeq G :\Longrightarrow TI(G') = TI(G)$ where $G', G \in \mathcal{G}$. Here, the relation $\simeq$ is identified with an isomorphism between G' and G. Namely, it is *a priori* valid that if two networks are *topologically identical* (i.e., *isomorphic*), then they also possess identical values of all topological indices. Although, the reverse correspondence is not universally true. This means that in general case $TI(G) = TI(G')$ does not imply that $G \simeq G'$. A topological index TI is said to be *complete* if the identity of $TI(G)$ and $TI(G')$ implies that the networks G and G' are isomorphic. On the other hand, a topological index TI is said to be *degenerate* if there exists at least two non-isomorphic networks G and G' such that $TI(G) = TI(G')$. Up to now, there is no known complete network invariant with respect to the class of all networks $\mathcal{G}$. This means that every topological index is degenerate to some extent. However, when we consider some subclass of $\mathcal{G}$, e.g., the set of all networks with n

nodes, then it is possible to define a complete topological index for this subclass. On the other hand, if $TI(G) \neq TI(G')$ for some topological index TI, then it can be firmly stated that the networks G and G' are not isomorphic [15, 16, 28, 43].

Topological indices (as well as vertex invariants) have many practical and theoretical applications in network data mining. For instance, they make the computation of isomorphism between two complex networks markedly easier. Recall that the so-called *network* (*graph*) *isomorphism problem* (i.e., the problem of deciding whether two given networks are topologically identical) is one of the classical topics of *Graph Theory* with an unknown computational complexity. Namely, for this problem there is no deterministic polynomial-time algorithm and simultaneously this problem has not been yet classified as NP-complete [32]. Nevertheless, there are several heuristics enabling to determnine whether any two networks are isomorphic or not. Unfortunately, these heuristics are time-consuming for large networks. Accordingly, in order to reduce the search space, some "preprocessing" methods can be applied to desicively classify certain pairs (or subsets) of networks as non-isomorphic. Such precursor steps are usually based on different topological indices or network centrality measures [32].

Note that apart from the network isomorphism problem, topological indices are ubiquitous in network data mining. For instance, in *Chemical Graph Theory* which is a subfield of mathematical chemistry dealing with network-theoretical aspects of chemical compounds or chemical reactions, topological indices are used in the so-called *quntitative structure-property relationships* (QSPR) or *quantitative structure-activity relationships* (QSAR) *studies* [43]. In such investigations, chemical molecules are modelled by the so-called *chemical graphs* (also known as *molecular graphs*) in which nodes correspond to atoms and edges correspond to chemical bonds (for instance, the network $G1$ in Figure 1 corresponds to methylcyclopropane). The fundamental idea of *Chemical Graph Theory* is that physicochemical properties and biological (pharmacological, toxicological) activities of diverse molecules can be investigated by using the information encoded in their corresponding chemical graphs. Consequently, the main purpose of all QSPR/QSAR studies is to relate the structure of a molecule to a defined quantitatively property and/or activity. This kind of methodology can be mathematically expressed by the following equation

$$\text{Property/Activity} = f\,(\text{molecular structure}) = f\,(\text{topological indices})\,.$$

Ideally, a good topological index should exhibit a low degree of degeneracy and a high degree of correlation with certain physicochemical properties (or biologi-

cal activities). Therefore, it is of paramount importance for many practical and theoretical applications to search for novel highly discriminating network invariants [43].

In network data mining, there are many known topological indices based on the distance matrix $D(G)$ [25, 43]. Also, it is widely recognized that these invariants can be easily computed using current computer techniques. The *Wiener index* $W(G)$ is the first distance-based network invariant introduced in 1947 by chemist Harold Wiener. This index is widely used in QSPR/QSAR studies. Its definition is as follows

$$W(G) := \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[D(G)\right]_{ij} = \frac{1}{2}\sum_{i=1}^{n} s(v_i).$$

In [37], M. Randić et al. introduced the so-called *natural Wiener index* $NW(G)$ whose formal definition can be expressed as follows

$$NW(G) := \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[ND(G)\right]_{ij} = \frac{1}{2}\sum_{i=1}^{n} s^{N}(v_i).$$

Here, we propose to introduce the *Jaccard-Wiener* and *cosine Wiener indices* (denoted by $JW(G)$ and $\cos W(G)$, respectively) whose formal definitions are given by

$$JW(G) := \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[JD(G)\right]_{ij} = \frac{1}{2}\sum_{i=1}^{n} s^{J}(v_i)$$

and

$$\cos W(G) := \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\cos D(G)\right]_{ij} = \frac{1}{2}\sum_{i=1}^{n} s^{\cos}(v_i).$$

One of the recently proposed distance-based topological invariant is the so-called *hyper-Wiener index* $WW(G)$ [43]. Its formula is expressed as follows

$$WW(G) := \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left[D(G)\right]_{ij}^{2} + \left[D(G)\right]_{ij}\right).$$

Note that the squared term, i.e, $\left[D(G)\right]_{ij}^{2}$ provides comparatively more weight to elongated networks. Consequently, the hyper-Wiener index should be well correlated with phenomena that are strongly dependent upon the topological size of a complex network. Here, we propose to extend the definition of $WW(G)$ to other distance matrices. Thus, we obtain three new network invariants, i.e.,

the *hyper-natural Wiener* ($NWW(G)$), *hyper-Jaccard-Wiener* ($JWW(G)$) and *hyper-cosine Wiener* ($\cos WW(G)$) *indices*. These topological invariants are defined as follows

$$NWW(G) := \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left([ND(G)]_{ij}^{2} + [ND(G)]_{ij}\right),$$

$$JWW(G) := \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left([JD(G)]_{ij}^{2} + [JD(G)]_{ij}\right)$$

and

$$\cos WW(G) := \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left([\cos D(G)]_{ij}^{2} + [\cos D(G)]_{ij}\right).$$

If $D(G)$ is the distance matrix connected with any complex network $G = (V(G), E(G))$ where $|V(G)| = n$, then its *Harary matrix* (also known as the *reciprocal distance matrix*) is a real symmetric $n \times n$ matrix, denoted by $RD(G)$, whose entries are given by the following condition $[RD(G)]_{ij} = \frac{1}{[D(G)]_{ij}}$ if $[D(G)]_{ij} \neq 0$ and $[RD(G)]_{ij} = 0$ otherwise [24,25,43]. Then, the so-called *Harary index* $H(G)$ associated with G is defined as follows

$$H(G) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[RD(G)]_{ij}.$$

Here, we generalize the above definition to other distance matrices. The *natural Harary matrix* (or *reciprocal natural distance matrix*), denoted by $RND(G)$, is given by the condition $[RND(G)]_{ij} = \frac{1}{[ND(G)]_{ij}}$ if $[ND(G)]_{ij} \neq 0$ and $[RND(G)]_{ij} = 0$ otherwise, the *Jaccard-Harary matrix* (or *reciprocal Jaccard distance matrix*), denoted by $RJD(G)$, is given by the condition $[RJD(G)]_{ij} = \frac{1}{[JD(G)]_{ij}}$ if $[JD(G)]_{ij} \neq 0$ and $[RJD(G)]_{ij} = 0$ otherwise and the *cosine Harary matrix* (or *reciprocal cosine distance matrix*), denoted by $R\cos D(G)$, is defined as follows $[R\cos D(G)]_{ij} = \frac{1}{[\cos D(G)]_{ij}}$ if $[\cos D(G)]_{ij} \neq 0$ and $[R\cos D(G)]_{ij} = 0$ otherwise. In analogy to the Harary index, we introduce the *natural*, *Jaccard* and *cosine Harary indices*. They are denoted by $NH(G)$, $JH(G)$ and $\cos H(D)$, respectively. Their formal definitions are identified with the subsequent expressions

$$NH(G) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[RND(G)]_{ij},$$

$$JH(G) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[RJD(G)\right]_{ij}$$

and

$$\cos H(G) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[R\cos D(G)\right]_{ij}.$$

One of the most discriminating topological indices is the *Balaban index* $J(G)$ [43]. For any complex network $G = (V(G), E(G))$ where $|V(G)| = n$, this invariant is given by the condition

$$J(G) := \frac{m}{\mu+1}\sum_{v_iv_j\in E(G)}\left(s(v_i)\,s(v_j)\right)^{-\frac{1}{2}}.$$

Here, $m = |V(E)|$ and μ is the so-called *cyclomatic number* of G, i.e., $\mu = m-n+c$ where c is the number of connected component included in G. Thus, μ is equal to the smallest number of edges which must be deleted from G such that no cycle remains. Note that the factor $\frac{m}{\mu+1}$ from the defining formula of the Balaban index brings about that the values of $J(G)$ does not necessarily rise with increasing number of nodes and cycles in G. In [37], M. Randić et al. introduced the "natural" analogue of $J(G)$ index, i.e, the *natual Balaban index*, denoted by $NJ(G)$, and expressed as follows

$$NJ(G) := \frac{m}{\mu+1}\sum_{v_iv_j\in E(G)}\left(s^N(v_i)\,s^N(v_j)\right)^{-\frac{1}{2}}.$$

It seems desirable to introduce the *Jaccard-Balaban* $(JJ(G))$ and *cosine Balaban* $(\cos J(G))$ *indices*. Their formulae are as follows

$$JJ(G) := \frac{m}{\mu+1}\sum_{v_iv_j\in E(G)}\left(s^J(v_i)\,s^J(v_j)\right)^{-\frac{1}{2}}$$

and

$$\cos J(G) := \frac{m}{\mu+1}\sum_{v_iv_j\in E(G)}\left(s^{\cos}(v_i)\,s^{\cos}(v_j)\right)^{-\frac{1}{2}}.$$

For a complex network $G = (V(G), E(G))$, H.P. Schultz et al. (cf. [43]) introduced the $PRS(G)$ index (or *Product of Row Sums index*) which is defined as the product of the distance sums $s(v_i)$

$$PRS(G) := \prod_{i=1}^{n} s(v_i) \text{ or } \log(PRS(G)) := \log\left(\prod_{i=1}^{n} s(v_i)\right) = \sum_{i=1}^{n} \log(s(v_i)),$$

where $\log(\cdot)$ denotes the natural logarithm. In almost all applications, the second expression (i.e., $\log(PRS(G))$) is advised due to the large values that can be achieved by the $PRS(G)$ index. Now, the generalization of the $PRS(G)$ index to other distance sums is straightforward. Thus, we obtained three novel topological indices, defined as the product of the natural, Jaccard and cosine distance sums, respectively. These new invariants are denoted by $PRS^N(G)$, $PRS^J(G)$ and $PRS^{\cos}(G)$, respectively. Their formulae are given by the following expressions

$$PRS^N(G) := \prod_{i=1}^{n} s^N(v_i) \text{ or } \log\left(PRS^N(G)\right) := \\ = \log\left(\prod_{i=1}^{n} s^N(v_i)\right) = \sum_{i=1}^{n} \log\left(s^N(v_i)\right),$$

$$PRS^J(G) := \prod_{i=1}^{n} s^J(v_i) \text{ or } \log\left(PRS^J(G)\right) := \\ = \log\left(\prod_{i=1}^{n} s^J(v_i)\right) = \sum_{i=1}^{n} \log\left(s^J(v_i)\right)$$

and

$$PRS^{\cos}(G) := \prod_{i=1}^{n} s^{\cos}(v_i) \text{ or } \log(PRS^{\cos}(G)) := \\ = \log\left(\prod_{i=1}^{n} s^{\cos}(v_i)\right) = \sum_{i=1}^{n} \log(s^{\cos}(v_i)).$$

As mentioned in Introduction, many complex networks are characterized by the occurence of hubs and the small world property. Consequently, D. Bonchev integrated the information on the network adjacencies and distances into single topological index [7,9]. The simplest manner to combine these two properties into

one quantity is to calculate the proportion of the network total adjacency to the network total distance. Thus, for any complex network $G = (V(G), E(G))$, it is possible to formulate the so-called *first Bourgas index* [7, 9], denoted by $B1(G)$, whose formal definition is as follows

$$B1(G) := \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[A(G)]_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}[D(G)]_{ij}}.$$

This topological index (also understood as the network complexity measure) rises with the increase of the cardinality of $E(G)$ and with the more compressed kind of topological arrangement. Therefore, the first Bourgas index can be used to quantitatively evaluate the "small-worldness" of a given complex network G. Three generalization of this index to other distance structures are given by the following conditions

$$NB1(G) := \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[A(G)]_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}[ND(G)]_{ij}},$$

$$JB1(G) := \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[A(G)]_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}[JD(G)]_{ij}}$$

and

$$\cos B1(G) := \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[A(G)]_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}[\cos D(G)]_{ij}}.$$

The invariants $NB1$, $JB1$ and $\cos B1$ are termed as the *first natural Bourgas*, *first Jaccard Bourgas* and *first cosine Bourgas indices*, respectively. On the other hand, the second Bourgas index $B2(G)$ (also introduced by D. Bonchev [7, 9]) which quantifies the "compactness" of a complex network $G = (V(G), E(G))$ where $|V(G)| = n$ is given by the following formula

$$B2(G) = \sum_{i=1}^{n}\frac{k_i}{s(v_i)}.$$

In this case, the *second natural Bourgas* ($NB2(G)$), *second Jaccard Bourgas* ($JB2(G)$) and *second cosine Bourgas* ($\cos B(G)$) *indices* are expressed by the subsequent formulae

$$NB2(G) := \sum_{i=1}^{n} \frac{k_i}{s^N(v_i)},$$

$$JB2(G) := \sum_{i=1}^{n} \frac{k_i}{s^J(v_i)}$$

and

$$\cos B2(G) := \sum_{i=1}^{n} \frac{k_i}{s^{\cos}(v_i)}.$$

In network data mining, three most popular *eigenvalue-based* topological indices derived from the distance matrix $D(G)$ where $G = (V(G), E(G))$ and $|V(G)| = n$ are the *distance spectral radius* of G, the *distance energy* of G and the *distance Estrada index* of G [19, 25, 43]. Note that for every undirected complex network G, the matrix $D(G)$ is real and symmetric. Therefore, the eigenvalues of $D(G)$ are also real and can be ordered in non-increasing order, $\rho_1 \geq \rho_2 ... \geq \rho_n$. Then, the distance spectral radius, denoted by $\rho(D(G))$, is given by the condition

$$\rho(D(G)) := \max\{|\rho_1|, |\rho_2|, ..., |\rho_n|\}$$

whereas the distance energy $DE(G)$ of G is defined as

$$DE(G) := \sum_{i=1}^{n} |\rho_i|.$$

In turn, the distance Estrada index, denoted by $DEE(G)$, for G is identified by the subsequent expression

$$DEE(G) := \sum_{i=1}^{n} e^{\rho_i}.$$

The eigenvalues of the matrices $ND(G)$, $JD(G)$ and $\cos D(G)$ are also real and can be ordered in non-increasing order. Denoting by ρ_i^N, ρ_i^J and $\rho_i^{\cos}$ the eigenvalues of the natural, Jaccard and cosine distance matrix, respectively, we obtain the following definitions

$$\rho(ND(G)) := \max\left\{\left|\rho_1^N\right|, \left|\rho_2^N\right|, ..., \left|\rho_n^N\right|\right\},$$

$$\rho\left(JD\left(G\right)\right) := \max\left\{\left|\rho_1^J\right|, \left|\rho_2^J\right|, ..., \left|\rho_n^J\right|\right\},$$

$$\rho\left(\cos D\left(G\right)\right) := \max\left\{\left|\rho_1^{\cos}\right|, \left|\rho_2^{\cos}\right|, ..., \left|\rho_n^{\cos}\right|\right\},$$

$$NDE\left(G\right) := \sum_{i=1}^{n}\left|\rho_i^N\right|,$$

$$JDE\left(G\right) := \sum_{i=1}^{n}\left|\rho_i^J\right|,$$

$$\cos DE\left(G\right) := \sum_{i=1}^{n}\left|\rho_i^{\cos}\right|,$$

$$NDEE\left(G\right) := \sum_{i=1}^{n}e^{\rho_i^N},$$

$$JDEE\left(G\right) := \sum_{i=1}^{n}e^{\rho_i^J}$$

and

$$\cos DEE\left(G\right) := \sum_{i=1}^{n}e^{\rho_i^{\cos}}.$$

Here, $\rho\left(ND\left(G\right)\right)$, $\rho\left(JD\left(G\right)\right)$ and $\rho\left(\cos D\left(G\right)\right)$ stand for the *natural*, *Jaccard* and *cosine distance spectral radius* of G, respectively, $NDE\left(G\right)$, $JDE\left(G\right)$ and $\cos DE\left(G\right)$ for the *natural*, *Jaccard* and *cosine distance energy* of G, respectively. In turn, $NDEE\left(G\right)$, $JDEE\left(G\right)$ and $\cos DEE\left(G\right)$ denote the *natural*, *Jaccard* and *cosine distance Estrada indices*, respectively.

The *distance polynomial*, denoted by $Ch\left(D\right)$, of any complex network $G = \left(V\left(G\right), E\left(G\right)\right)$ where $\left|V\left(G\right)\right| = n$ is identified with the characteristic polynomial of its distance matrix $D\left(G\right)$ [43], i.e.,

$$Ch\left(D\right) = \det\left(x\mathbf{I} - D\right) = \sum_{k=0}^{n}c_k x^{n-k},$$

where $\mathbf{I}$ is the identity matrix and c_k are the coefficients of this polynomial. It seems justifiable to generalize the above definition to other distance matrices. Thus, the *natural*, *Jaccard* and *cosine distance polynomials* of G (denoted by $Ch\left(ND\right)$, $Ch\left(JD\right)$ and $Ch(\cos D)$, respectively) are given by the following formulae

$$Ch\left(ND\right) = \det\left(x\mathbf{I} - ND\right) = \sum_{k=0}^{n}c_k^N x^{n-k},$$

$$Ch\left(JD\right)=\det\left(x\mathbf{I}-JD\right)=\sum_{k=0}^{n}c_{k}^{J}x^{n-k}$$

and

$$Ch\left(\cos D\right)=\det\left(x\mathbf{I}-\cos D\right)=\sum_{k=0}^{n}c_{k}^{\cos}x^{n-k},$$

where c_k^N, c_k^J and $c_k^{\cos}$ are the coefficients of the natural, Jaccard and cosine distance polynomials, respectively. One of the characteristic polynomial-base topological indices is the *Hosoya Z' index* [43]. Namely, for any complex network $G=\left(V\left(G\right),E\left(G\right)\right)$ where $\left|V\left(G\right)\right|=n$, this invariant is defined as follows

$$Z'\left(G\right)=\sum_{k=0}^{n}\left|c_{k}\right|,$$

where $|c_k|$ are the absolute values of the coefficients of the distance polynomial associated with G. In order to generalize this index to the newly proposed distance polynomials, let us introduce the following definitions

$$NZ'\left(G\right)=\sum_{k=0}^{n}\left|c_{k}^{N}\right|,$$

$$JZ'\left(G\right)=\sum_{k=0}^{n}\left|c_{k}^{J}\right|$$

and

$$\cos Z'\left(G\right)=\sum_{k=0}^{n}\left|c_{k}^{\cos}\right|,$$

where $\left|c_k^N\right|$, $\left|c_k^J\right|$ and $|c_k^{\cos}|$ are the absolute values of the coefficients of the natural, Jaccard and cosine distance polynomials. In this context, the quantities NZ', JZ' and $\cos Z'$ are termed as the *natural*, *Jaccard* and *cosine Hosoya Z' indices*.

Also, some information theory-based complexity measure for a complex network $G=\left(V\left(G\right),E\left(G\right)\right)$ where $\left|V\left(G\right)\right|=n$ was proposed by E.V. Konstantinova et al. In [27], the *vertex complexity index* (denoted by $H_D\left(v_i\right)$) for any vertex $v_i\in V\left(G\right)$ was introduced and defined as the entropy of its shortest distances from all other vertices in G, i.e.,

$$H_{D}\left(v_{i}\right):=-\sum_{j=1}^{n}\frac{d_{G}\left(v_{i},v_{j}\right)}{s\left(v_{i}\right)}\log_{2}\left(\frac{d_{G}\left(v_{i},v_{j}\right)}{s\left(v_{i}\right)}\right).$$

Then, the global measure has the form

$$H_D^n(G) := \sum_{i=1}^{n} H_D(v_i).$$

Here, H_D^n denotes the *information distance index* of G. Denoting by $H_N(v_i)$, $H_J(v_i)$ and $H_{\cos}(v_i)$ the entropies of the natural, Jaccard and cosine distances for $v_i \in V(G)$, we obtain the following expressions

$$H_N(v_i) := -\sum_{j=1}^{n} \frac{d_G^N(v_i,v_j)}{s^N(v_i)} \log_2\left(\frac{d_G^N(v_i,v_j)}{s^N(v_i)}\right),$$

$$H_N^n(G) := \sum_{i=1}^{n} H_N(v_i),$$

$$H_J(v_i) := -\sum_{j=1}^{n} \frac{d_G^J(v_i,v_j)}{s^J(v_i)} \log_2\left(\frac{d_G^J(v_i,v_j)}{s^J(v_i)}\right),$$

$$H_J^n(G) := \sum_{i=1}^{n} H_J(v_i),$$

$$H_{\cos}(v_i) := -\sum_{j=1}^{n} \frac{d_G^{\cos}(v_i,v_j)}{s^{\cos}(v_i)} \log_2\left(\frac{d_G^{\cos}(v_i,v_j)}{s^{\cos}(v_i)}\right)$$

and

$$H_{\cos}^n(G) := \sum_{i=1}^{n} H_{\cos}(v_i).$$

The quantities H_N^n, H_J^n and $H_{\cos}^n$ stand for the *information natural distance*, *information Jaccard distance* and *information cosine distance indices*, respectively.

In Section 7, it will be demonstrated numerically that most of the newly identified topological invariants have a significantly reduced level of degeneracy.

6. Datasets and computational methods

All exemplary complex networks used in the present study are publically available. Four of the six networks used in this study are classified as spatial (they are denoted by ptn1, ptn2, ptn3 and ptn4) while the other two are considered as social (they are denoted by sn1 and sn2). The ptn1 network is the largest

connected component of the graph *USairports* from [12]. This dataset is the network of passenger flights between airports in the United States in 2010 December. We removed all loops and multiple edges from this largest component. All edges were treated as undirected. The ptn2 network is the dataset describing all air connections between US states during 2014 (from the Bureau of Transportation Statistics) downloaded from [20]. The ptn3 network is the largest connected component of the graph of London metro downloaded from [38]. The ptn4 network is the dataset of the Goettingen bus connections [31]. In our study, we treat all edges of this network as undirected. The sn1 network is the dataset of frequent associations between 62 dolphins [29]. This network was downloaded from [5]. The sn2 network is the dataset *UKfaculty* from [12]. This dataset is the personal friendsip network of a faculty of a UK university. In the present work, we treat all edges of this dataset as undirected and unweighted. Thus, we used in our study six real-world networks (four public transportation and two social networks). The basic statistical characteristics of these six network datasets are contained in Table 2.

Table 2
The statistical parameters of six real-world complex networks

Index	ptn1	ptn2	ptn3	ptn4	sn1	sn2
$\|V(G)\|$	745	53	318	257	62	81
$\|E(G)\|$	4618	1150	366	328	159	577
$dens(G)$	0.0167	0.8345	0.0073	0.0100	0.0841	0.1781
$L(\lambda_2)$	0.0743	7.9652	0.0065	0.0085	0.1730	1.3261
$Q(w)$	0.3368	0.0182	0.7605	0.8046	0.4888	0.4317
$Q(fg)$	0.4310	0.0298	0.8249	0.8157	0.4955	0.4442
$Q(le)$	0.4096	0.0315	0.7922	0.7523	0.4912	0.3970
$\langle l \rangle$	3.4472	1.1655	13.8241	11.8418	3.3570	2.0975

For a complex network $G = (V(G), E(G))$ where $V(G) = n$ and $E(G) = m$, its *density* ($dens(G)$) is calculated according to the equation $dens(G) = \frac{2m}{n(n-1)}$. The *algebraic connectivity* (denoted by $L(\lambda_2)$) of G is identified with the second smallest eigenvalue of the Laplacian matrix $L(G)$ [19]. The *Laplacian matrix* is defined as $L(G) = Deg(G) - A(G)$ where $Deg(G)$ is the degree matrix for G and $A(G)$ is its adjacency matrix. The *degree matrix* is a $n \times n$ diagonal matrix whose entries are defined as follows $[Deg(G)]_{ij} = k_i$ if $v_i = v_j$ and $[Deg(G)]_{ij} = 0$ otherwise. The *modularity* Q of a complex network with respect to some division of its vertex set $V(G)$ is computed according to the subsequent equation $Q = \frac{1}{2m} \sum_{v_i, v_j} \left([A(G)]_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$ where c_i is the type of v_i, c_j that of v_j,

Table 3
The statistical parameters of two network models

Index	rand.ptn2	rand.sn2
$\|V(G)\|$	53	81
$\|E(G)\|$	1150	577
$dens(G)$	0.8345	0.1781
$L(\lambda_2)$	35.5083	5.8150
$Q(w)$	0.0168	0.1761
$Q(fg)$	0.0291	0.1872
$Q(le)$	0.0291	0.1704
$\langle l \rangle$	1.1655	1.8842

the summation is carried out over all v_i and v_j pairs of nodes and $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise [11]. In our study, we calculated the modularity of six real-world networks (and two randomly generated models) subjected to three different types of division of $V(G)$. These divisions are identified with the following algorithms: the *walktrap community finding algorithm* ($Q(w)$) [35], the *fast greedy modularity optimalization algorithm* ($Q(fg)$) and the *leading eigenvector method* ($Q(le)$) [11]. The *average path length* $\langle l \rangle$ of G is calculated according to the equation $\langle l \rangle = \frac{1}{n(n-1)} \sum_{v_i \neq v_j} d_G(v_i, v_j)$. Two random networks, denoted by rand.ptn2 and rand.sn2, were generated according to the $G(n, m)$ *model* where n is the number of nodes and m is the number of edges [19,33]. These edges are chosen uniformly randomly from the collection of all possible edges. Eight statistical properties of these two models are summarized in Table 3. The results from this Table as well as from Tables 6 and 9 are averages based on 200 simulation trials.

All discriminating tests were carried out on a synthetic dataset of all exhaustively generated non-isomorphic, undirected and connected networks having up to 7 nodes with the exception of the unique network with $|V(G)| = 1$ and $|E(G)| = 0$. This dataset, denoted by $\mathbf{G}$, contains 995 small graphs (1 graph with $|V(G)| = 2$, 2 graphs with $|V(G)| = 3$, 6 graphs with $|V(G)| = 4$, 21 graphs $|V(G)| = 5$, 112 graphs with $|V(G)| = 6$ and 853 graphs with $|V(G)| = 7$). Note that these quantities are in agreement with the Pólya enumeration theory [34]. All networks from $\mathbf{G}$ are numbered from 1 (the network K_2) to 995 (the network K_7). Here, the symbol K_n denotes the complete network with n nodes. In order to quantify the *uniqueness* (i.e., the *degree of degeneracy*)of a particular topological index TI

or centrality measure C, the *sensitivity index* $S(TI/C)$ indtroduced by E.V. Konstantinova was used [27]. This index is defined as

$$S(TI/C) = \frac{|\mathbf{G}| - |ndv(\mathbf{G})|}{|\mathbf{G}|}$$

where $|\mathbf{G}|$ denotes the cardinality of a dataset $\mathbf{G}$ on which TI or C were tested (in our case $|\mathbf{G}| = 995$) and $|ndv(\mathbf{G})|$ stands for the number of degeneracies of TI or C within $\mathbf{G}$. It is immediately apparent that when $S(TI/C) = 1$, then the analyzed dataset $\mathbf{G}$ does not contain any pair (or subset) of non-isomorphic networks with the same value of TI or with the same vector of centralities. Also, it can be easily demonstrated that the sensitivity index $S(TI/C)$ is dependent on the selected decimal places. Consequently, in discriminating tests, all topological indices and all centrality measures were calculated with an accuracy of 6 decimal places.

In the present paper, all linear relationships between two variables were assessed by the *Pearson correlation coefficient* r.

All simulations and computations included in the present work were conducted in the R programming language [13, 18, 22, 40]. The source codes of several R functions used in this paper are published in [45].

7. Numerical results and discussion

7.1. Correlations between the closeness-type and harmonic-type centrality measures

In this Section, we will explore the linear correlations between four closeness-type and four harmonic-type centrality measures defined on six real-world complex networks and two random network models (cf. Table 4, 5, 6). J.R.F. Ronqui and G. Travieso observed that complex networks can be characterized by some peculiar pattern of linear correlations between centrality measurements [42]. They called this phenomena as the *centrality correlation profile* of a given network. From Tables 4 and 5, it can be seen that the correlation profiles of the ptn1, ptn3, ptn4, sn1 and sn2 datasets with respect to the closeness-type and harmonic-type centralities are comparable. Namely, it can be spotted that two centralities based on the natural distance, i.e, NC and NHC are (within the closeness-type or harmonic-type, respectively) negatively correlated with all other centrality measures based

on the geodesic, Jaccard and cosine distances. On the other hand, all closeness-type and all harmonic-type centralities based on the geodesic, Jaccard and cosine distances are (within the closeness-type or harmonic-type, respectively) positively correlated. The measurements carried out on the artificial model rand.sn2 confirm this regularity. In turn, the correlation profile of the ptn2 network is different. In this case, all closeness type and all harmonic-type centralities are (within the closeness-type or harmonic-type, respectively) very strongly positively correlated (in the first type r is always greater than 0.97 and in the second type r is above 0.94). Also, all closeness-type and all harmonic-type measures evaluated on the rand.ptn2 model are (within the closeness-type or harmonic-type, respectively) very strongly positively correlated. From Table 2, it can be seen that the ptn2 network is characterized (among all analyzed datasets) by its significantly higher density (0.8345) and algebraic connectivity (7.9652). Also, the ptn2 network differs from other studied datasets by its lower modularity in all three types of measurements (i.e., $Q(w) = 0.0182$, $Q(fg) = 0.0298$ and $Q(le) = 0.0315$). Note that the random model of this network, i.e., the rand.ptn2 model is also marked by its high algebraic connectivity (35.5083) and its low modularity ($Q(w) = 0.0168$, $Q(fg) = 0.0291$ and $Q(le) = 0.0291$).

Therefore, it can be hypothetized that the correlation profiles of complex networks with respect to four closeness-type and four harmonic-type centrality measures is determined by the density, algebraic connectivity and modularity of these highly structured relational entities.

Table 4
The Pearson correlation coefficients between four closeness-type centrality measures defined on six real-world complex networks

Centralities	ptn1	ptn2	ptn3	ptn4	sn1	sn2
CC/NC	−0.6975	0.9723	−0.4805	−0.3827	−0.7277	−0.8509
CC/JC	0.8096	0.9832	0.5716	0.4575	0.5359	0.8432
$CC/\cos C$	0.8513	0.9827	0.6489	0.5171	0.5850	0.9065
NC/JC	−0.7156	0.9979	−0.2962	−0.3187	−0.6008	−0.7437
$NC/\cos C$	−0.7271	0.9985	−0.4259	−0.4358	−0.6300	−0.7708
$JC/\cos C$	0.9914	0.9998	0.9766	0.9735	0.9847	0.9841

Table 5
The Pearson correlation coefficients between four harmonic-type centrality measures defined on six real-world complex networks

Centralities	ptn1	ptn2	ptn3	ptn4	sn1	sn2
HC/NHC	−0.7189	0.9466	−0.5840	−0.5311	−0.8847	−0.9254
HC/JHC	0.8299	0.9459	0.5943	0.3731	0.7161	0.8366
$HC/\cos HC$	0.8731	0.9491	0.6378	0.4554	0.6781	0.8749
NHC/JHC	−0.7733	0.9994	−0.3411	−0.3494	−0.7415	−0.7389
$NHC/\cos HC$	−0.7317	0.9995	−0.3593	−0.3812	−0.6680	−0.7651
$JHC/\cos HC$	0.9862	0.9999	0.9801	0.9758	0.9739	0.9930

Table 6
The Pearson correlation coefficients between four closeness-type and four harmonic-type centrality measures defined on two network models

Centralities	rand.ptn2	rand.sn2	Centralities	rand.ptn2	rand.sn2
CC/NC	0.9868	−0.9357	HC/NHC	0.9868	−0.9726
CC/JC	0.9920	0.8990	HC/JHC	0.9920	0.9002
$CC/\cos C$	0.9911	0.8991	$HC/\cos HC$	0.9911	0.8984
NC/JC	0.9993	−0.8266	NHC/JHC	0.9993	−0.8200
$NC/\cos C$	0.9993	−0.8281	$NHC/\cos HC$	0.9993	−0.8179
$JC/\cos C$	0.9999	0.9971	$JHC/\cos HC$	0.9999	0.9973

7.2. Compexity of six real-world networks

In Section 4, we have introduced eight partition-dependent entropy measures (i.e., the so-called mean information content indices) as well as eight partition-independent entropy measures (i.e., the so-called structural information content indices). Table 7 presents the values of mean information content invariants evaluated on six real-world complex networks. In the case of all networks, it can be seen that the mean information content corresponding to three closeness-type centralities (NC, JC and $\cos C$) and three harmonic-type centralities (NHC, JHC and $\cos HC$) are pairwise equal. On the other hand, the mean information content evaluated with respect to the "classical" closeness centrality and "classical" harmonic centrality attains the same value only in the case of the ptn2 network. In all other cases, the values of $\overline{I}_{CC}$ and $\overline{I}_{HC}$ are different. Also, the measurements of this invariant performed on the artificial model rand.ptn2 produced the identical

values of $\overline{I}_{CC}$ and $\overline{I}_{HC}$ (cf. Table 9). These same measurements carried out on the rand.sn2 model gave different results. Also, it was observed that in the cases of the datasets ptn1, ptn2, sn1 and sn2 the mean information content indices evaluated with respect to the newly proposed centralities, i.e., NC, JC, $\cos C$, NHC, JHC and $\cos HC$ are equal. This same phenomenon can be seen in the case of measurements done on the rand.ptn2 and rand.sn2 models. In the cases of the datasets ptn1, ptn2, sn1 and sn2, the mean information content invariants assessed with respect to "non-classical" centralities have the higher values than their analogues based on "classical" closeness and "classical" harmonic centralities. Also, from Table 7, it can be observed that the mean information content of the ptn2 network evaluated with respect to all eight centrality measures is the smallest among all analyzed datasets. Therefore, it can be speculated that the partition-dependent entropy measures derived from four closeness-type and four harmonic-type centrality measures enable to differentiate real-world complex networks with respect to their correlation profiles.

Table 8 summarizes the values of structural information content indices evaluated with respect to four closeness-type and four harmonic-type centrality measures defined on six real-world complex networks. From this table, it can be observed that the partition-independent entropies of the dataset ptn2 assessed with respect to all eight centralities are the smallest among all studied networks. Also, it can be infer from Table 8 that the values of structural information content indices within a given dataset are very similar but not identical (with a few exceptions). The standard deviation of the values of I_C where $C \in \{CC, NC, JC, \cos C, HC, NHC, JHC, \cos HC\}$ evaluated on six real-world networks ranges from 0.005 (the sn2 dataset) to 0.0245 (the ptn1 dataset). In two artificial models rand.ptn2 and rand.sn2, the standard deviation for the measurement of I_C is equal to 0.0008 and 0.0007, respectively (cf. Table 9). Also, it can be spotted that in the case of ptn1, ptn3, ptn4, sn1 and sn2 datasets, the values of I_C where $C \in \{JC, \cos C, JHC, \cos HC\}$ are *slightly* higher than the values of this invariant evaluated with respect to other centrality measures. In the case of ptn2 network, the reverse relationships is observed. The measurements performed on two artificial models rand.ptn2 and rand.sn2 confirmed this regularity.

In summary, it can be stated that the structural information content indices can be used in order to differentiate complex networks with respect to their correlation profiles.

Table 7
The partition-dependent entropy measures defined on six real-world complex networks evaluated with respect to four closeness-type and four harmonic-type centralities

Entropy	ptn1	ptn2	ptn3	ptn4	sn1	sn2
$\overline{I}_{CC}$	8.6268	4.1713	8.2123	7.9045	5.2725	5.5618
$\overline{I}_{NC}$	9.2485	5.3294	5.8985	6.8870	5.8897	6.3399
$\overline{I}_{JC}$	9.2485	5.3294	5.8585	6.8484	5.8897	6.3399
$\overline{I}_{\cos C}$	9.2485	5.3294	5.8881	6.9025	5.8897	6.3399
$\overline{I}_{HC}$	9.0022	4.1713	8.2814	7.9434	5.7606	5.9602
$\overline{I}_{NHC}$	9.2485	5.3294	5.8985	6.8870	5.8897	6.3399
$\overline{I}_{JHC}$	9.2485	5.3294	5.8585	6.8484	5.8897	6.3399
$\overline{I}_{\cos HC}$	9.2485	5.3294	5.8881	6.9025	5.8897	6.3399

Table 8
The partition-independent entropy measures defined on six real-world complex networks evaluated with respect to four closeness-type and four harmonic-type centralities

Entropy	ptn1	ptn2	ptn3	ptn4	sn1	sn2
I_{CC}	9.5205	5.7164	8.2716	7.9657	5.9334	6.3274
I_{NC}	9.4965	5.7104	8.3073	7.9999	5.9405	6.3333
I_{JC}	9.5410	5.7064	8.3129	8.0056	5.9541	6.3393
$I_{\cos C}$	9.5407	5.7022	8.3129	8.0056	5.9539	6.3383
I_{HC}	9.5134	5.7210	8.2603	7.9635	5.9314	6.3257
I_{NHC}	9.4701	5.7006	8.3070	7.9996	5.9370	6.3321
I_{JHC}	9.5410	5.6950	8.3129	8.0056	5.9540	6.3391
$I_{\cos HC}$	9.5406	5.6884	8.3129	8.0056	5.9537	6.3376

7.3. Discriminating tests

As mentioned in Section 5, the so-called discriminatory power is one of the fundamental characteristics of any network invariant TI [43]. This property measures its capability to distinguish among the non-isomorphic networks. Many studies in network data mining are devoted to quantitative assessments of the degree of degeneracy of diverse topological indices. For instance, D. Bonchev and N. Trinajstić

evaluated the discriminatory abilities of information and topological invariants between 45 alkane trees [10]. C. Raychaudhury et al. also conducted their studies on 45 alkane trees as well as on 19 monocyclic networks [39]. E.V. Konstantinova et al. investigated the discriminating abilities of different invariants on 1443032 polycyclic networks and 3473141 network trees [27]. Also, M. Dehmer et al. quantified the discriminatory power of many topological indices on several datasets of all exhaustively generated networks possessing n nodes [15, 16].

Table 9
The partition-dependent and partition independent entropy measures defined on two network models evaluated with respect to four closeness-type and four harmonic-type centralities

Entropy	rand.ptn2	rand.sn2	Entropy	rand.ptn2	rand.sn2
$\overline{I}_{CC}$	3.2112	4.4058	I_{CC}	5.7266	6.3384
$\overline{I}_{NC}$	5.7279	6.3399	I_{NC}	5.7261	6.3383
$\overline{I}_{JC}$	5.7279	6.3399	I_{JC}	5.7257	6.3398
$\overline{I}_{\cos C}$	5.7279	6.3399	$I_{\cos C}$	5.7249	6.3397
$\overline{I}_{HC}$	3.2112	5.0986	I_{HC}	5.7274	6.3382
$\overline{I}_{NHC}$	5.7279	6.3399	I_{NHC}	5.7260	6.3382
$\overline{I}_{JHC}$	5.7279	6.3399	I_{JHC}	5.7256	6.3398
$\overline{I}_{\cos HC}$	5.7279	6.3399	$I_{\cos HC}$	5.7247	6.3397

Here, we present our results evaluating the discriminatory abilities of the newly introduced centralities and invariants carried out on the dataset of all exhaustively generated small networks having up to 7 vertices. Table 4 includes the results of twelve discriminating experiments performed on the dataset $\mathbf{G}$. In these tests, twelve centrality measures C were evaluated on all networks from $\mathbf{G}$.

Table 10
The sensitivity index (S) of twelve centrality measures (C)

C	$S(C)$	C	$S(C)$	C	$S(C)$
DC	0.1709	CC	0.4804	HC	0.5116
BC	0.9397	NC	0.9146	NHC	0.9146
EC	0.9196	JC	0.9940	JHC	0.9940
PRC	0.9508	$\cos C$	0.9940	$\cos HC$	0.9940

Two networks from $\mathbf{G}$ are said to be indistinguishable with respect to C if they possess the same vector (in the non-decreasing order) of this centrality. In this case, the measure C is said to be degenerate. From Table 4, it can be seen that all tested centrality measures are degenerate to some extent. The degree centrality DC exhibits the lower level of uniqueness with respect to the dataset $\mathbf{G}$. On the other hand, four newly introduced centralities, i.e., JC, $\cos C$, JHC and $\cos HC$ are extremely sensitive with respect to $\mathbf{G}$. These four measures have only six degeneracies within the dataset $\mathbf{G}$. Thus, in the case of two closeness-type centralities based on the Jaccard and cosine distances, the improvement in the specificity is equal to 98.84 % in comparison with the classical closeness centrality. In turn, in the case of two harmonic-type centralities also based on these newly proposed distances, the improvement in the specificity is equal to 98.77 % compared to the classical harmonic centrality. The degrees of degeneracy of the eigenvector EC, natural closeness NC and natural harmonic NHC centralities are comparable.

Consequently, it can be uttered that calculations of the Jaccard and cosine closeness centralities as well as the Jaccard and cosine harmonic centralities can be used as some precursor steps in the network isomorphism problem in order to categorically classify certain pairs (or subsets) of networks as non-isomorphic.

Table 5 contains the degrees of degeneracy of the newly introduced topological indices as well as their analogues based on the shortest path distance evaluated on the dataset $\mathbf{G}$.

Table 11
The sensitivity index (S) of fourty eight TIs derived from the geodesic, natural, Jaccard and cosine distance matrices

TI	$S(TI)$	TI	$S(TI)$	TI	$S(TI)$	TI	$S(TI)$
W	0.0111	WW	0.0221	H	0.0432	J	0.8302
NW	0.6633	NWW	0.6633	NH	0.6633	NJ	0.9940
JW	0.9779	JWW	0.9779	JH	0.9759	JJ	0.9920
$\cos W$	0.9759	$\cos WW$	0.9779	$\cos H$	0.9759	$\cos J$	0.9920
PRS	0.4714	$B1$	0.0312	$B2$	0.4432	$\rho(D)$	0.9226
$PRS(ND)$	0.9146	$NB1$	0.8261	$NB2$	0.9839	$\rho(ND)$	0.9025
$PRS(JD)$	0.9940	$JB1$	0.9658	$JB2$	0.9839	$\rho(JD)$	0.9799
$PRS(\cos D)$	0.9940	$\cos B1$	0.9719	$\cos B2$	0.9839	$\rho(\cos D)$	0.9799
DE	0.9397	DEE	0.9779	Z'	0.5126	H_D^n	0.5116
NDE	0.9025	$NDEE$	0.9126	NZ'	0.9106	H_N^n	0.9146
JDE	0.9799	$JDEE$	0.9940	JZ'	0.9899	H_J^n	0.9920
$\cos DE$	0.9799	$\cos DEE$	0.9940	$\cos Z'$	0.9899	$H_{\cos}^n$	0.9940

From this table, some regularities can be deduced. Namely, in the cases of topological indices based on the distance matrix (i.e., the Wiener-type, hyper-Wiener-type, Harary-type, Balaban-type and product of row sums-type invariants), on the distance and adjacency matrix (i.e., the first and second Bourgas-type invariants), on the characteristic polynomial (i.e., the Hosoya Z'-type invariants) and on Information Theory (i.e., the information distance-type invariants), it can be observed that the indices derived from the natural, Jaccard and cosine distance matrices are overwhelmingly more specific with respect to the dataset **G** than their "shortest path distance" analogues. On the other hand, in the cases of eigenvalue-based indices (i.e., distance spectral radius-type, distance energy-type and distance Estrada-type invariants), only indices derived from the Jaccard and cosine distance structures are significantly more unique with respect to **G** than their "geodesic distance" counterparts. The eigenvalue-based indices derived from the natural distance matrix are more degenerate than their "classical" analogues. Thus, in the cases of nine topological indices (i.e., NJ, JJ, $\cos J$, $PRS\,(JD)$, $PRS\,(\cos D)$, $JDEE$, $\cos DEE$, H_J^n and $H_{\cos}^n$), the sensitivity index is above 0.99. Hence, these invariants are extremely specific with respect to the dataset **G**. In the cases of seventeen topological indices (i.e., JW, $\cos W$, JWW, $\cos WW$, JH, $\cos H$, $JB1$, $\cos B1$, $NB2$, $JB2$, $\cos B2$, $\rho\,(JD)$, $\rho\,(\cos D)$, JDE, $\cos DE$, JZ' and $\cos Z'$), the sensitivity index is greater than 0.96. Hence, these invariants can be regarded as strongly unique with regard to **G**.

In summary, it can be asserted that the topological indices derived from the natural, Jaccard and cosine matrices can be helpful in network data mining and their computations can be used (for instance) in order to desicively classify some complex networks as non-isomorphic.

Table 12

The sensitivity index (S) of eight partition-dependent and eight partition-independent entropy measures evaluated with respect to closeness-type and harmonic-type centralities

TI	$S\,(TI)$	TI	$S\,(TI)$	TI	$S\,(TI)$	TI	$S\,(TI)$
$\overline{I}_{CC}$	0.002	$\overline{I}_{HC}$	0.002	I_{CC}	0.4573	I_{HC}	0.4905
$\overline{I}_{NC}$	0.002	$\overline{I}_{NHC}$	0.002	I_{NC}	0.8422	I_{NHC}	0.8452
$\overline{I}_{JC}$	0.003	$\overline{I}_{JHC}$	0.003	I_{JC}	0.8030	I_{JHC}	0.8794
$\overline{I}_{\cos C}$	0.002	$\overline{I}_{\cos HC}$	0.002	$I_{\cos C}$	0.8824	$I_{\cos HC}$	0.9176

Table 12 presents the degrees of degeneracy of the mean information content-type and the structural information content-type indices derived from four closeness-type and four harmonic-type centrality measures. All these invariants were evaluated on the dataset **G**. From this Table, it can be observed that $\overline{I}_{CC}$, $\overline{I}_{NC}$, $\overline{I}_{JC}$, $\overline{I}_{\cos C}$, $\overline{I}_{HC}$, $\overline{I}_{NHC}$, $\overline{I}_{JHC}$ and $\overline{I}_{\cos HC}$ complexity measures are highly degenerate. They uniquely identify only two or three networks within the dataset **G**. On the other hand, eight partition-independent complexity measures are more specific with respect to this dataset. Six structural information content indices derived from six newly introduced centralities, i.e., I_{NC}, I_{JC}, $I_{\cos C}$, I_{NHC}, I_{JHC} and $I_{\cos C}$ are less degenerate than their counterparts derived from the "classical" closeness and "classical" harmonic centralities, i.e., I_{CC} and I_{HC}.

7.4. The newly defined topological indices as cyclicity measures

In this Subsection, we will apply the concepts introduced in Section 5 to the class of all connected cyclic graphs with five vertices in order to study their cyclicity. This octet of small networks is given in Figure 2.

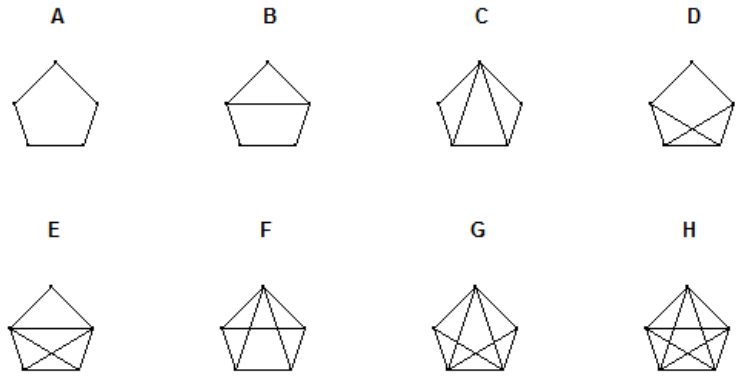

Fig. 2. All connected cyclic graphs with five vertices

Molecular cyclicity is a structural property that has not yet been strictly defined. Nevertheless, from an intuitive point of view, the notion of cyclicity is conceptually linked to the number of cycles (i.e., closed paths) formed from a fixed number of nodes. Consequently, it can be observed that the higher quantity of closed paths, the higer cyclicity of a network. This notion is mainly studied in Chemical Graph Theory.

The set of all connected cyclic graphs with five vertices was studied in [2,30,36]. The topological indices from Section 5 evaluated on these eight small networks produced the orders presented in Table 13.

Table 13
The orders of eight graphs from Figure 2 induced by fourty eight TIs

entry	Order	TI
I	$A < B < C < D < E < F < G < H$	JJ, $\cos J$, $NB1$, $JB1$, $\cos B1$, $NB2$, $JB2$, $\cos B2$
II	$H < G < F < E < D < C < B < A$	NW, JW, $\cos W$, JWW, $PRS(ND)$, $PRS(JD)$, $PRS(\cos D)$, $\rho(D)$, $\rho(ND)$, $\rho(JD)$, $\rho(\cos D)$, NDE, JDE, $\cos DE$, DEE, $NDEE$, $JDEE$, $\cos DEE$, $\cos Z'$
III	$A < B < D < C < F < E < G < H$	J, $B2$, H^n_D
IV	$H < G < E < F < C < D < B < A$	NWW, $\cos WW$, PRS
V	$A < B < C = D < E = F < G < H$	H, $B1$
VI	$H < G < F = E < D = C < B < A$	W, WW
VII	$H < G < F < E < C < D < B < A$	DE
$VIII$	$H < G < F < E = D < C < B < A$	Z'
IX	$B < A < C < D < E < F < G < H$	NJ
X	$F < G < D < A < B < C < E < H$	NH
XI	$F < A < D < B < G < C < E < H$	JH, $\cos H$
XII	$F < G < D < H < E < C < B < A$	NZ'
$XIII$	$F < G < H < D < E < C < B < A$	JZ'
XIV	$F < D < G < B < C < E < A < H$	H^n_N, H^n_J, $H^n_{\cos}$

From this table, it can be observed that eight topological indices (i.e, JJ, $\cos J$, $NB1$, $JB1$, $\cos B1$, $NB2$, $JB2$, $\cos B2$ and $\cos B2$) ordered eight cyclic networks illustrated in Figure 2 according to their increased cyclicity. This means that the lowest values are ascribed to the cycle C_5 and the highest values to the complete network K_5. The same order of this dataset was induced by the Wiener sum and resistance Balaban indices in [2] as well as by the number of spanning trees invariant in [30]. On the other hand, nineteen indices (NW, JW, $\cos W$, JWW, $PRS(ND)$, $PRS(JD)$, $PRS(\cos D)$, $\rho(D)$, $\rho(ND)$, $\rho(JD)$, $\rho(\cos D)$, NDE, JDE, $\cos DE$, DEE, $NDEE$, $JDEE$, $\cos DEE$ and $\cos Z'$) produced the reverse order by ascribing the highest values to the cycle C_5 and the smallest values to the complete graph K_5. Note that the same order was produced by the Kirchhoff and Kirchhoff sum indices in [2]. The above results are also in agreement with Randić's order of this octet of small networks. In [36], he used the Wiener sum indices (obtained from the quotient matrix D/Δ, where Δ is the detour matrix) as cyclicity measures.

The indices J, $B2$ and H_D^n also ascribed the smallest values to the cycle C_5 and the highest values to the complete graph K_5. In addition, these invariants produced two reversals (i.e., $C \leftrightarrow D$ and $E \leftrightarrow F$). In turn, the indices NWW, $\cos WW$ and PRS induced the reverse order, that is, the smallest values are possessed by the K_5 network and the highest values by the C_5 network. These invariants also produced two reversals (i.e., $D \leftrightarrow C$ and $F \leftrightarrow E$). The indices H and $B1$ also ascribed the smallest values to the cycle C_5 and the highest values to the complete graph K_5. These invariants equally valuate the graphs C and D as well as the graphs E and F. Note that the order induced by the indices H and $B1$ is identical to the order given by the vertex-degree sums (or equivalently, twice the number of edges, i.e., $2m$) and by the cyclomatic number μ. On the other hand, the indices W and WW produced the reverse order. These invariants also equally valuate the graphs C and D as well as E and F. The distance energy ascribed the lowest value to the complete graph K_5 and the highest value to the cycle C_5. This invariant induced one reversal (i.e., $D \leftrightarrow C$). The Hosoya index Z' also ascribed the lowest value to the network K_5 and the highest value to the network C_5. This invariant equally valuates the graphs D and E.

To summarize the above results, it can be asserted that the orders I, III and V from Table 13 are intuitively acceptable. Hence, the topological indices inducing these orders can be used as cyclicity measures in the field of network data mining. In turn, the reverse orders (i.e., the orders II, IV, VI, VII and $VIII$) can be simply changed to the correct orders by adding a minus sign to the relevant topological indices. Therefore, the invariants inducing the reverse orders can also be regarded as cyclicity measures. On the other hand, the indices NJ, NH, JH, $\cos H$, NZ', JZ', H_N^n, H_J^n and $H_{\cos}^n$ (similarly as the reciprocal spanning-tree densities in [30]) produced an intuitively unacceptable orders and they can not be applied as cyclicity measures.

All detailed numerical values of fourty eight topological indices evaluated on the set of all connected cyclic graphs with five vertices are published in [45].

7.5. The newly defined centrality measures in QSPR studies

It can be easily observed that if $C = \{x_1, x_2, ..., x_n\}$ is any vector of centrality measures then the quantities $AM(C)$ (i.e., the arithmetic mean of C), $GM(C)$ (i.e., the geometric mean of C), $HM(C)$ (i.e., the harmonic mean of C) and $\|C\|_p$ (i.e., the $p-$norm of C) can be regarded as topological indices. Similarly as two entropymeasures $\overline{I}_C$ and I_C, the quantities $AM(C)$, $GM(C)$, $HM(C)$ and

$\|C\|_p$ are network-level invariants derived from C. In this Subsection, we will show numerically that many of these invariants can be used in QSPR studies of saturated alkanes.

Saturated alkanes are considered as an especially attractive class of organic compounds which are often used as a starting point for any QSPR studies. One of the approaches often taken in such investigations is to choose a certain class of alkanes (for instance, C_8, C_9 or C_{10} isomers) in order to obtain comparable results and to avoid the so-called size effect. In the present work, we have used the dataset of octane isomers. This benchmark dataset consists of 18 octane isomers and contain 16 physicochemical properties of these compounds. International Academy of Mathematical Chemistry advised to use this reference dataset for any preliminary evaluation of modelling capabilities of newly proposed topological indices. This dataset can be downloaded from [23]. For our study, we singled out the subsequent properties of octanes: the boiling point (BP), the enthalpy of formation (HFORM), the enthalpy of vaporization (HVAP) and the standard enthalpy of vaporization (DHVAP). The rationale for selecting these properties is that for this collection of physicochemical parameters at least one of the tested invariants exhibits a relatively good linear correlation (i.e., $|r| > 0.8$).

Table 14
The Pearson correlation coefficients between four physicochemical properties of octanes and the central tendency-type topological indices derived from four closeness-type centrality measures; the values above |0.8| are in bold

TI	BP	HFORM	HVAP	DHVAP
$AM(CC)$	−0.4911	−0.479	−0.7084	−0.7975
$AM(NC)$	**−0.8493**	**−0.8833**	**−0.9135**	**−0.9262**
$AM(JC)$	**−0.8511**	**−0.883**	**−0.8821**	**−0.8843**
$AM(\cos C)$	**−0.8599**	**−0.8895**	**−0.9084**	**−0.9141**
$GM(CC)$	−0.5041	−0.4928	−0.7196	**−0.8069**
$GM(NC)$	**−0.8398**	**−0.8755**	**−0.9106**	**−0.9276**
$GM(JC)$	**−0.8457**	**−0.8792**	**−0.8802**	**−0.8857**
$GM(\cos C)$	**−0.8495**	**−0.8831**	**−0.9039**	**−0.9142**
$HM(CC)$	−0.5163	−0.5059	−0.73	**−0.8158**
$HM(NC)$	**−0.8271**	**−0.8659**	**−0.9076**	**−0.9294**
$HM(JC)$	**−0.8393**	**−0.8745**	**−0.8774**	**−0.8861**
$HM(\cos C)$	**−0.8363**	**−0.8746**	**−0.8971**	**−0.9122**

From Table 14, it can be observed that all central tendency-type topological indices derived from the newly introduced centrality measures are satisfactorily linearly correlated with four properties of octanes. Also, all central tendency-type invariants derived from the natural, Jaccard and cosine centralities exhibit significantly higher correlations than their counterparts derived from the "classical" closeness centrality.

On the other hand, our initial studies indicated that in almost all cases the central tendency-type topological indices derived from the harmonic-type centralities are not satisfactorily correlated with this set of properties (data not shown).

In this paper, the $p-$norm-type indices were computed for $p = 2$ and $p = 3$. From Table 15, it can be spotted that all $p-$norm-type network invariants derived from the newly introduced centrality measures exhibit $|r| > 0.8$. In all cases, they are significantly better correlated with four physicochemical properties of octanes than their counterparts based on "classical" closeness centrality.

On the other hand, the $p-$norm-type topological indices derived from the harmonic-type centralities are weakly linearly correlated with this set of physicochemical properties (data not shown).

Table 15
The Pearson correlation coefficients between four physicochemical properties of octanes and the p-norm-type topological indices derived from four closeness-type centrality measures; the values above |0.8| are in bold

TI	BP	HFORM	HVAP	DHVAP
$\|CC\|_2$	−0.4779	−0.4654	−0.6972	−0.788
$\|NC\|_2$	**−0.8567**	**−0.8901**	**−0.9164**	**−0.9251**
$\|JC\|_2$	**−0.8554**	**−0.8859**	**−0.8832**	**−0.8821**
$\|\cos C\|_2$	**−0.8675**	**−0.8941**	**−0.9109**	**−0.9125**
$\|CC\|_3$	−0.4648	−0.4525	−0.6864	−0.7788
$\|NC\|_3$	**−0.8626**	**−0.8957**	**−0.919**	**−0.9239**
$\|JC\|_3$	**−0.8586**	**−0.8879**	**−0.8834**	**−0.8792**
$\|\cos C\|_3$	**−0.873**	**−0.8972**	**−0.9119**	**−0.9097**

From Table 16, it can be observed that all partition-dependent entropy measures evaluated with respect to the newly defined centralities are significantly better linearly correlated with BP, HFORM, HVAP and DHVAP than their analogue based on the "classical" closeness centrality. Also, from this Table, it can be seen that only one invariant (i.e., I_{NC}) is satisfactorily linearly correlated with the

above properties of octanes. In turn, in almost all cases, the partition-independent entropy measures derived from the newly introduced centralities exhibit higher correlations with these properties than their analogue defined with respect to "classical" closeness centrality. Only two indices (i.e., I_{JHC} and $I_{\cos HC}$) are satisfactorily correlated with this set of physicochemical properties of octanes.

The partition-dependent entropy measures evaluated with respect to the closeness-type or harmonic-type centralities are weakly correlated with this set of parameters.

Table 16
The Pearson correlation coefficients between four physicochemical properties of octanes and the partition-independent entropy measures evaluated with respect to four closeness-type and four harmonic-type centralities; the values above |0.8| are in bold

TI	BP	HFORM	HVAP	DHVAP
I_{CC}	−0.2669	−0.3183	−0.0786	0.0445
I_{NC}	**0.8938**	**0.9184**	**0.9115**	**0.8828**
I_{JC}	0.7803	0.7863	0.7554	0.7033
$I_{\cos C}$	0.7864	0.7605	0.7395	0.6749
I_{HC}	0.3846	0.327	0.5857	0.6799
I_{NHC}	0.6846	0.5531	0.6044	0.507
I_{JHC}	**0.8394**	**0.8186**	**0.8425**	**0.8142**
$I_{\cos HC}$	**0.8938**	**0.8315**	**0.8895**	**0.8574**

To sum up these results, it can be uttered that the newly introduced centralities and the network-level topological indices derived from these measures exhibit better modelling abilities than their "classical" analogues. Consequently, it can be speculated that these new invariants can be used as molecular descriptors (i.e., structurally meaningful numbers) in QSPR studies.

8. Conclusions

In this report, we have introduced two novel distance structures, i.e., the Jaccard and cosine distance matrices. Both these two-dimensional arrays satisfy Euclid's postulates. From these matrices and the natural distance matrix introduced by Randić et al. [37], we have derived six novel centrality measures as well as

several novel topological indices. In the experiments carried out on six real-world complex networks and two random models, we have determined that linear correlations between four closeness-type and four harmonic-type measures are mainly governed by the density, algebraic connectivity and modularity of the underlying networks. We have also determined that the so-called correlation profile of a complex network depends on its mean and structural information content. In discriminating tests performed on the dataset of all axhaustively generated small networks, we have established that the newly introduced centralities and topological indices in almost all cases are more sensitive with respect to this dataset than their "classical" analogues.

The last two Subsections have presented some applications of the concepts introduced in this work to Chemical Graph Theory. Namely, it has been shown that most of the newly introduced topological indices can be used as cyclicity measures. Molecular cyclicity is a parameter that influence the behavior of molecules. As stated in [8] "[...] *molecular cyclicity is expressed by a number of topological patterns or structural transformations that allow one to compare the cyclic complexity of structures*". Consequently, it is possible to single out different aspects of molecular cyclicity. Undoubtedly, eight intuitively acceptable orders from Table 13 reveal diverse facets of this notion. In our opinion, it seems justifiable to assert that the novel cyclicity measures proposed in this work will be very helpful in elucidating different levels of molecular intricacy.

The second important practical achievement of this paper was to demonstrate that the newly introduced network invariants can be used as molecular descriptors. Recall that "[t]*he molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*" [43]. It is widely recognized that any molecular descriptor must be invariant with regard to labelling and numbering of atoms and any spatial translations of molecules. Also, any molecular descriptor must be defined by a computable mathematical expression and its values must be in an acceptable numerical range. A *good* molecular descriptor should also be correlated with at least one experimentally measurable physicochemical property and/or biological (pharmacological, toxicological) activity. In Subsection 7.5, it has been demonstrated that several invariants derived from the newly proposed centrality measures are satisfactorily correlated with four selected physicochemical properties of octanes. In almost all cases, they have higher correlation abilities than their counterparts based on the "classical" concepts of centrality. Therefore, it seems justifiable to utter that these novel notions are able to extract some pieces of chemical information

from molecular graphs and can be regarded as "good" molecular descriptors. It can be hypothesized that invariants derived from the natural, Jaccard and cosine distance matrices will be used in molecular modelling.

Acknowledgment

The author would like to express his thanks to the anonymous reviewer for his valuable comments and suggestions.

References

1. Albert R., Barabási A.-L.: *Statistical mechanics of complex networks*. Rev. Mod. Phys. **74** (2002), 47–97.
2. Babić D., Klein D.J., Lukovits I., Nikolić S., Trinajstić N.: *Resistance-distance matrix: a computational algorithm and its application.* Int. J. Quantum Chem. **90** (2002), 166–176.
3. Barabási A.-L., Albert R.: *Emergence of scaling in random networks*. Science **286** (1999), 509–512.
4. Barthélemy M.: *Spatial networks*. Phys. Rep. **499** (2011), 1–101.
5. Batagelj V.: vlado.fmf.uni-lj.si/pub/networks/data/bio/dolphins.net.
6. Bavelas A.: *Communication patterns in task-oriented groups*. J. Acoust. Soc. Am. **22** (1950), 725–730.
7. Bonchev D.: *A simple integrated approach to network complexity and node centrality*. In: Analysis of Complex Networks. From Biology to Linguistics. Dehmer M., Emmert-Streib F. (eds.), Wiley-VCH, Weinheim 2009, 47–53.
8. Bonchev D., Balaban A.T., Liu X., Klein D.J.: *Molecular cyclicity and centricity of polycyclic graphs. I. Cyclicity based on resistance distances and reciprocal distances*. Int. J. Quantum Chem. **50** (1994), 1–20.
9. Bonchev D., Buck G.A.: *Quantitative measures of network complexity*. In: Complexity in Chemistry, Biology, and Ecology. Bonchev D., Rouvray D.H. (eds.), Springer, New York 2005, 191–235.
10. Bonchev D., Trinajstić N.: *Information theory, distance matrix and molecular branching*. J. Chem. Phys. **67** (1977), 4517–4533.
11. Clauset A., Newman M.E.J., Moore C.: *Finding community structure in very large networks*. Phys. Rev. E **70** (2004), 066111.

12. Csardi G.: *igraphdata: A collection of network data sets for the 'igraph' package*. R package version 1.0.1, https://CRAN.R-project.org/package=igraphdata.
13. Csardi G., Nepusz T.: *The igraph software package for complex network research*. InterJournal Complex Syst. **1695** (2006), http://igraph.org.
14. Dehmer M.: *Information theory of networks*. Symmetry **3** (2011), 767–779.
15. Dehmer M., Grabner M., Furtula B.: *Structural discrimination of networks by using distance, degree and eigenvalue-based measures*. PLoS ONE **7** (2012), e38564.
16. Dehmer M., Grabner M., Varmuza K.: *Information indices with high discriminative power for graphs*. PloS ONE **7** (2012), e31214.
17. Dorogovtsev S.N., Mendes J.F.F.: *Evolution of Networks: From Biological Networks to the Internet and WWW*. Oxford University Press, Oxford 2003.
18. Dray S., Dufour A.B.: *The ade4 package: implementing the duality diagram for ecologists*. J. Stat. Soft. **22** (2007), 1–20.
19. Estrada E.: *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, Oxford 2011.
20. Eubank N.: http://github.com/nickeubank/gis_in_r/blob/master/RGIS6_Data/.
21. Gower J.C., Legendre P.: *Metric and Euclidean properties of dissimilarity coefficients*. J. Classification **3** (1986), 5–48.
22. Hausser J., Strimmer K.: *Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks*. J. Mach. Learn. Res. 10 (2009), 1469–1484.
23. International Academy of Mathematical Chemistry: www.moleculardescriptors.eu.
24. Ivanciuc O., Balaban T.S., Balaban A.T.: *Design of topological indices. Part 4*. Reciprocal distance matrix, related local vertex invariants and topological indices.* J. Math. Chem. **12** (1993), 309–318.
25. Janežič D., Miličević A., Nikolić S., Trinajstić N.: *Graph Theoretical Matrices in Chemistry*, University of Kragujevac, Kragujevac 2007.
26. Klein D.J.: *Graph geometry, graph metrics and Wiener*. MATCH Commun. Math. Comput. Chem. **35** (1997), 7–27.
27. Konstantinova E.V., Skorobogatov V.A., Vidyuk M.V.: *Applications of information theory in chemical graph theory*. Indian J. Chem. Sect. A **42** (2003), 1227–1240.

28. Koschűtzki D., Lehman K.A., Peeters L., Richter S., Tenfelde-Podehl D., Zlotowski O.: *Centrality indices*. In: Network Analysis: Methodological Foundations. Brandes U., Erlebach T. (eds.), Springer, Berlin 2005.
29. Lusseau D., Schneider K., Boisseau O.J., Haase P., Slooten E., Dawson S.M.: *The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations*. Behav. Ecol. sociobiol. **54** (2003), 396–405.
30. Mallion R.B., Trinajstić N.: *Reciprocal spanning-tree density: a new index for characterising the intricacy of a (poly)cyclic molecular-graphs*. MATCH Commun. Math. Comput. Chem. **48** (2003), 97–116.
31. Manitz J.: *NetOrigin: Origin estimation for propagation processes on complex networks*. R package version 1.0-2, https://CRAN.R-project.org/package=NetOrigin.
32. Meghanathan N.: *Exploiting the discriminating power of eigenvector centrality measure to detect graph isomorphism*. IJFCST **5** (2015), 1–13.
33. Newman M.E.J.: *The structure and function of complex networks*. SIAM Review **45** (2003), 167–256.
34. Pólya G., Read R.C.: *Combinatorial Enumeration of Groups, Graphs and Chemical Compounds*, Springer-Verlag, New York 1987.
35. Pons P., Latapy M.: *Computing communities in large networks using random walks*. J. Graph Algorithms Appl. **10** (2006), 191–218.
36. Randić M.: *On characterization of cyclic structures*. J. Chem. Inf. Comput. Sci. **37** (1997), 1063–1071.
37. Randić M., Pisanski T., Novič M., Plavšić D.: *Novel graph distance matrix*. J. Comput. Chem. **31** (2010), 1832–1841.
38. Rappaport D.: research.cs.queensu.ca/˜daver/235/C1352963146/E20070302133910.
39. Raychaudhury C., Ray S.K., Ghosh J.J., Roy A.B., Basak S.C.: *Discrimination of isomeric structures using information theoretic topological indices*. J. Comput. Chem. **5** (1984), 581–588.
40. R Core Team: *R: a language and environment for statistical computing*. R foundation for Statistical Computing, Vienna 2015, http://www.R-project.org/.
41. Rochat Y.: *Closeness Centrality Extended to Unconnected Graphs: the Harmonic Centrality Index*, Application of Social Networks Analysis, Zurich 2009.
42. Ronqui J.R.F., Travieso G.: *Analyzing complex networks through correlations in centrality measurements*. J. Stat. Mech.: Theory Exp. **2015** (2015), P05030.

43. Todeschini R., Consonni V.: *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, Weinheim 2009.
44. Watts D.J., Strogatz S.H.: *Collective dynamics of 'small-world' networks*. Nature **393** (1998), 440–442.
45. Wilczek P.: https://github.com/PiotrWilczeknet/R-functions-1.