Silesian J. Pure Appl. Math.
vol. 9, is. 1 (2019), 9–47

Piotr WILCZEK

Computer Laboratory, Poznań, Poland

IDENTIFYING INFLUENTIAL NODES IN THE GENOME-SCALE METABOLIC NETWORKS

Abstract. The present article introduces two novel centrality indices which can be used in order to characterize the genome-scale metabolic networks. The deliberate attack simulation experiments conducted on two Barabási-Albert models and four genome-scale metabolic networks demonstrate that the proposed ranking methods are effective in identifying essential nodes in complex networks. Also, the Principal Component Analysis reveals that the *Kendall centrality correlation profile* can be used to describe the metabolic networks and distinguish them from their random counterparts with the preserved degree distribution.

1. Introduction

Complex systems are composed of connected elements whose interactions are, in substance, *nonlinear*. Instances of complex systems encompass our society, the internet, our brain and cellular interactions. The modelling of complex systems has attracted attention of many scholars from several different branches, e.g., mathematics, physics, biology, computer science just to mention a few. The representation of complex interacting systems as *complex networks* is commonplace in modern science and engineering. In this setting, *nodes* of a complex network

Mathematics Subject Classification: 05C12, 92C40.
Keywords: complex networks, centrality measures, attack simulations.
Corresponding author: P.Wilczek (piotr.wilczek.net@onet.pl).
Received: 02.07.2019.

correspond to constituents of a complex system and *edges* of a network are formal representatives of interactions within a complex system [10, 27]. While such an approach in terms of vertices and edges is almost universal, the systems so described are highly diverse and composite. This fact stemmed from the phenomena that vertices and edges in real-world complex networks are tailored to perform different and inequivalent functions. Therefore, many formal metrics were developed in order to quantitatively assess the heterogeneity and intricacy of empirical networks. For instance, the so-called *network centrality measures* try to quantify the relative importance of nodes (or edges) within a complex network [10, 25, 27]. Consequently, a centrality is a key property of complex networks that has effect on the behavior of dynamic processes occurring on a network, like synchronization or epidemic spreading and can provide essential information about the organization and functions of these structures.

Note that the accumulation of the results in *network theory* is parallel to the expansion of new technologies and the growth of computer literacy in different branches of science. For instance, it can be observed that current innovations in advanced molecular techniques have incited the development of databases that systematically store knowledge of how different biological entities interact. Such structured data can be naturally represented by complex networks where vertices correspond to biological objects (e.g., genes, proteins, transcription factors and metabolites) and edges correspond to interactions between them. Roughly speaking, the molecular networks can be categorized into four main groups, i.e., the *protein-protein interaction networks* (denoted by PPI), the *regulatory networks*, the *signal transduction networks* and the *metabolic networks* [10]. These four categories share important structural properties with other real-world networks in different fields ranging from the internet to social systems. It was also demonstrated that topological analyses carried out on molecular networks can be regarded as a valuable guide to understand and identify factors that play a major role in the underlying biological processes. For instance, PPI networks possess a small number of highly connected protein vertices (known as *hubs*) and many poorly connected nodes (see [49] and the references cited therein). The empirical as well as computational studies demonstrated that the deletion of a hub protein is more likely to be lethal than the deletion of a non-hub protein. This regularity is referred to as the *centrality-lethality rule* [20]. In general terms, the above-mentioned principle indicates that there exists a high correlation between the essentiality of a protein and its topological centrality in a PPI network. Accordingly, in current proteomics, network centrality measures have been an important method for identifying essential proteins (see [36, 49] and the references cited therein).

Metabolic networks are the best-studied molecular networks. These complex systems are composed of metabolites and biochemical reactions converting these metabolites into each other. In a metabolic network, vertices correspond to metabolites and edges correspond to biochemical reactions between them [19]. In contrast to other molecular networks such as signal transduction networks or regulatory networks, complete topologies of metabolic networks can be easily gained from annotated genomes. Namely, an exponentially increasing number of organisms possess sequenced genomes with many determined encoded proteins. Supposing that the annotated enzymes are expressed, it is possible to reconstruct the metabolic network of the organism. Thus, the so-called *genome-scale metabolic networks* are identified with manually curated models that reflect our comprehension of the metabolic processes occurring in a living organism [28]. These networks can be regarded as an indispensable tool to obtain biological knowledge from metabolomic data. The metabolic networks which have been reconstructed from genome information enable in-depth mechanistic interpretations through metabolic flux simulations and network analysis. For instance, by graph-theoretical analyses of the topology of genome-scale metabolic networks, it is possible to conjecture about the functionality of metabolism.

In [19], H. Jeong and coworkers noted that the connectivity of metabolic networks follows approximately a *power law*, i.e., the frequency, denoted by $P(k)$, of metabolites participating in k biochemical reactions is given by the expression $P(k) \sim ak^{-\gamma}$, where a is a constant and γ is a positive exponent. The above fact is tantamount to the situation in which most metabolites are involved in only few reactions and some metabolites (known as *hub metabolites*) are involved in many reactions. These hub metabolites are more frequent then would be expected, for instance, in random networks. This *scale-free property* of metabolic networks is responsible for their robustness in the sense that these networks often remain intact when a large portion of randomly selected vertices is deleted from the network. However, if a small portion of the hub metabolites of the network is deleted the network is likely to become disintegrated into several separate clusters [2,7,8,28].

In several papers, metabolites were ranked based on their position within a metabolic network in order to determine their relative significance (see [19,26,46] and the references cited therein). In [46], S. Wuchty and P.F. Stadler argued that, in the case of metabolic networks, the meaning of a centrality measure is obvious. Namely, according to the above authors, central metabolites are the crossroads of the network and are historically oldest ones. They maintain that, in metabolic networks, the notion of a centrality measure simultaneously reflect the importance as well as the age of a metabolite. Therefore, in metabolic networks, in

addition to establishing the significance of a metabolite for metabolism, centrality indices can be very useful in determining the core of metabolism by computational tools. Thereby, the notion of the metabolic core can be identified with a central, highly connected part of metabolism which provides main substrates for many anabolic reactions and is probably evolutionary highly conserved. It should also be mentioned that, in current metabolomics, the ranking of metabolites in a metabolic network can be used to identify *drug targets* for the next generation of medicines [31]. Consequently, it can be claimed that the development of effective ranking methods applicable to metabolic networks is of paramount importance from both theoretical as well as practical perspectives.

Complex networks and centrality measures have been at the forefront of statistical mechanics for more than a decade. This branch of physics uncovered several basic rules that govern complex networks and can be applied to a wide range of complex systems, from the internet to social and biological networks. Evaluating the importance of nodes in complex networks can undoubtedly help to understand them and to develop the capability to control them.

In this paper, we propose two new ranking methods which can be thought of as extensions of the measures considered in [45]. To test their effectiveness in identifying essential nodes in the genome-scale metabolic networks, we follow the methodology borrowed from statistical mechanics. Namely, the main part of the present work consists of the targeted attack simulation experiments in which two newly introduced centrality measures are juxtaposed with the degree centrality as well as with three representative state-of-the-art centrality indices and the efficacy of the attack strategies based on these metrics is comparatively evaluated. Also, the Principal Component Analysis is employed in order to explore the rank correlations between the above centrality measures defined on the genome-scale metabolic networks.

The rest of the paper is organized as follows: Section 2 presents the necessary background on *graph theory* and recent centrality indices which are used as contrast metrics in Section 5. Section 3 proposes two novel ranking methods whereas Section 4 describes the methodology and datasets used in Section 5. The results and discussion are contained in Section 5. Finally, Section 6 concludes the paper.

2. Background

In the present work, it is assumed that all considered complex networks are modelled by simple graphs of the general form $G = (V(G), E(G))$ where $V(G) = \{v_1, v_2, \ldots, v_n\}$ is the vertex set and $|E(G)| = m$ is the edge set. For two vertices $v_i, v_j \in V(G)$, $v_i v_j$ means that v_i and v_j are *adjacent*, i.e., $v_i v_j \in E(G)$. The *neighborhood* of the node $v_i \in V(G)$, denoted by $N_G(v_i)$, is identified with the following set: $N_G(v_i) = \{w \in V(G) \mid wv_i \in E(G)\}$. The symbol k_i refers to the *degree* of the vertex $v_i \in V(G)$. Undoubtedly, $k_i = |N_G(v_i)|$. The *degree centrality* corresponding to the node $v_i \in V(G)$, denoted by $DC(v_i)$, is given by $DC(v_i) = \frac{k_i}{n-1}$ or simply by $DC(v_i) = k_i$. A complex network with the vertex set $|V(G)| = n$ can be represented by the *adjacency matrix* $A(G) \in \{0,1\}^{n \times n}$ whose entries are given by the term $a_{ij} = 1$ if there is an edge between v_i and v_j and $a_{ij} = 0$ otherwise [10, 27]. The *shortest path* (*geodesic*, *topological*) *distance* between two vertices $v_i, v_j \in V(G)$, denoted by d_{ij}, is identified with the number of edges in any shortest path connecting them [10, 27].

As previously mentioned, the notion of a centrality measure has many interpretations and for most of them it is possible to indicate a lengthy catalogue of different proposed metrics. Here, we briefly review three recently suggested centrality indices which are used in our intensional attack simulation experiments as benchmark measures.

In [6], D. Chen et al. proposed the *local centrality measure* (also known as the *semi-local centrality measure*), denoted by SL, which can be perceived as a compromise between the low-relevant and highly degenerate degree centrality and other time-consuming indices. This centrality takes into account the nearest and the next nearest neighbors of a given node. Formally, the local centrality measure of the vertex $v_i \in V(G)$ is defined by the subsequent condition

$$Q(v_j) = \sum_{v_n \in N_G(v_j)} No(v_n),$$
$$SL(v_i) = \sum_{v_j \in N_G(v_i)} Q(v_j),$$

where $No(v_n)$ is the number of the nearest and the next nearest neighbors of the node v_n. Using the SIR model, the above authors demonstrated that the SL measure can effectively identify important vertices. Namely, D. Chen and coworkers showed that, in comparison with other widely used centrality measures, the proposed SL index performs much better than the degree and betweenness

centralities and almost as good as the closeness centrality while with much lower computational complexity [6].

In turn, in [11], L. Fei et al. introduced the centrality measure whose definition is inspired by the *inverse square law*. Recall that the inverse square law is any physical law stating that a given physical quantity or intensity is inversely proportional to the square of the distance from the source of that physical quantity. The above researchers postulate that the mutual attraction between two vertices in a complex network follows the inverse square law. They defined this mutual attraction, denoted by F, by the subsequent formula

$$F\left(v_i, v_j\right) = \frac{k_i \times k_j}{d_{ij}^2}.$$

Then, they argued that the sum of the attraction of a vertex to all other nodes in the network can be understood as a property of the vertex itself. This sum can be identified with the intensity of the node and can be perceived as its importance. Thus, the intensity of the node $v_i \in V\left(G\right)$ in the complex network $G = \left(V\left(G\right), E\left(G\right)\right)$ is given by the following condition

$$I\left(v_i\right) = \sum_{\substack{j=1 \\ j \neq i}}^{n} F\left(v_i, v_j\right).$$

Naturally, the intensity of a vertex can be interpreted as its centrality. This measure is also denoted by I. In their experiments [11], the above authors demonstrated that the newly proposed index based on the inverse square law can effectively identify influential nodes in complex networks.

In [47], Xu et al. proposed a *multiattribute centrality measure* based on the k-shell index and the structural holes. Recall that the *k-shell index*, denoted by KS, is based on the notion of the *k-shell decomposition* of a complex network. The k-shell decomposition allows us to identify the core and the periphery of the network. This procedure relies on the subsequent steps [23]:

a) delete all vertices of the degree $k_i = 1$ and also all their links. This step may reduce the degree of other nodes to 1;

b) delete vertices whose degree has been reduced to 1 and all their edges until all the remaining vertices have the degree greater than 1. All of the removed nodes and the edges between them are considered to form the k-shell with $KS = 1$;

c) delete vertices with the degree $k_i = 2$ and all their edges in the remaining network until all of the remaining vertices have the degree greater than 2. Thus, the newly deleted nodes and the edges between them are regarded to form the k-shell with $KS = 2$;

d) repeat the analogous process for higher values of KS.

Consequently, at the end of the above presented decomposition, each node is associated with its own KS index which indicates the topological location of this vertex within the network. Kitsak and coworkers demonstrated that the highly connected vertices may possess notably different KS indices and may be situated either in the core or in the periphery of the network [23]. These researchers also found that the degree of a vertex is not necessarily correlated with its spreading capability whereas the KS index can be a better predictor of the spreading influence of that node [23]. Unfortunately, the k-shell decomposition algorithm classifies many vertices with different degree into the same k-shell and, consequently, the resulting ranking list has too many ties.

In order to overcome the above indicated limitation of the k-shell index, many KS-based indices have been proposed. The proposal of Xu and coworkers depends on the notion of the so-called *links diversity assessment index*, denoted by V, which is expressed by the following formula [47]:

$$V(v_i) = \sum_{KS=1}^{KS_{\max}} \frac{l_i^{KS}}{N_{KS}} \times e_i^{KS},$$

where

$$e_i^{KS} = \frac{KS(v_i)}{KS_{\max}}.$$

In this setting, $KS_{\max}$ denotes the maximum value of the KS index, l_i^{KS} represents the number of edges from the node v_i to the other nodes with the k-shell value equal to KS, N_{KS} refers to the number of nodes with the k-shell value equal to KS and e_i^{KS} is the normalized KS value of the node v_i. Thus, the V index takes into account the location of nodes as well as their interactions with neighboring vertices at different layers. The second concept on which the proposal of Hu et al. depends is the notion of the *structural holes*. The theory of the structural holes is developed in the works of R.S. Burt who introduced the so-called *network constraint coefficient*, denoted by S, to quantify the structural holes of a complex

networks [5]. This index can be defined by the following expressions

$$s\left(v_i, v_j\right) = \left(q_{ij} + \sum_m q_{im} q_{mj}\right)^2, \qquad i \neq j \neq m,$$

$$S\left(v_i\right) = \sum_{v_j \in N_G(v_i)} s\left(v_i, v_j\right)$$

and

$$q_{ij} = \frac{a_{ij}}{\sum\limits_{v_j \in N_G(v_i)} a_{ij}},$$

where q_{ij} refers to the weight proportion of the node v_j in all the adjacency of the node v_i, the node m is the common neighbor of v_i and v_j and the quantity $\sum\limits_m q_{im} q_{mj}$ depends on the number of common neighbors m of v_i and v_j. Then, the new multiattribute centrality measure proposed by Hu and coworkers, denoted by VKC, is identified with the subsequent formula [47]:

$$VKC\left(v_i\right) = \sum_{v_j \in N_G(v_i)} V\left(v_j\right) \times \frac{k_i}{\sum\limits_{v_j \in N_G(v_i)} k_j} \times \left(\frac{1}{S\left(v_i\right)}\right)^{\alpha},$$

where α is a tunable parameter whose range lies in the interval $[1, \langle k \rangle]$, where $\langle k \rangle$ is the average degree of the network. In the present work, α is set to one.

In this section, we have shortly presented three modern centrality measures: the local centrality measure (SL), the centrality measure based on the inverse square law (I) and the multiattribute centrality measure based on the k-shell index and the structural holes (VKC). These indices (with the classical degree centrality) are used as contrast measures in Section 5. Note that the definitions of these indices encompass the information on the nearest and the next nearest neighbors of a focal node (the SL measure), on the degree and the topological distances between nodes (the I measure) and on the KS index, the degree and the network constraint index of a focal vertex (the VKC measure). Consequently, it can can be claimed that the above selected centrality metrics are representative for contemporary *network science*.

3. The proposed methods

One of the widely used centrality measure is the *closeness centrality*. This classical centrality index quantifies the importance of a vertex in any complex network $G = (V(G), E(G))$ as the inverse of the sum of the distances to other nodes in G. However, it was noticed that this centrality metric has two main drawbacks. Namely, the closeness centrality measure is well-defined only for connected networks and, even if the network is connected, its values are dominated by distant vertices. In order to surmount these defects, M. Marchiori and V. Latora [30] and (independently) Y. Rochat [37] (cf. also [17]) introduced the harmonic mean of the geodesic distances between a given node and all other nodes in a complex network $G = (V(G), E(G))$ as a centrality index. The resulting measure – the so-called (*shortest path*) *harmonic centrality*, denoted by HC, is given by the following expression

$$HC(v_i) = \sum_{\substack{v_j \in V(G) \\ v_j \neq v_i}} \frac{1}{d_{ij}}.$$

In [45], the present author generalized the above concept and proposed the *natural harmonic centrality measure*, denoted by NHC. For a complex network $G = (V(G), E(G))$, this newly defined index has the form

$$NHC(v_i) = \sum_{\substack{v_j \in V(G) \\ v_j \neq v_i}} \frac{1}{d_{ij}^N}.$$

In the above expression, the term d_{ij}^N refers to the *natural distance* between two vertices $v_i, v_j \in V(G)$. Recall that, for a complex network $G = (V(G), E(G))$, the notions of the natural distance and the *natural distance matrix*, denoted by $ND(G)$, were proposed by M. Randić and coworkers [34]. They observed that for any complex network $G = (V(G), E(G))$, where $|V(G)| = n$, it is possible to treat the rows of its adjacency matrix $A(G)$ as points in the $n-$dimensional Euclidean space. Then, the natural distance between two vertices $v_i, v_j \in V(G)$ is given by the following condition [34]:

$$d_{ij}^N = \Big\{ \sum_{i=1}^{n} (a_{ik} - a_{jk})^2 \Big\}^{\frac{1}{2}}.$$

Consequently, the elements of the natural distance matrix, denoted by nd_{ij}, are given by the natural distances between the points corresponding to nodes of G

in the $n-$dimensional Euclidean space, i.e., $nd_{ij} = d_{ij}^N$ if $v_i \neq v_j$ and $nd_{ij} = 0$ if $v_i = v_j$. Note that the natural distance between nodes in a complex network is, in general, a *pseudo-distance* as it may be equal to zero for two different vertices [34]. This will hapen whenever the corresponding adjacency rows are identical. This is the case for pairs of nonadjacent nodes which possess the same neighbors. C.D. Godsil and W.L. Kocay named such nodes as *psudosimilar* [12]. Accordingly, for the natural distance, the condition that a metric is positive (i.e., generally, $d(x,y) > 0$ if $x \neq y$ where d is any metric function) is relaxed to the requirement that the metric is non-negative (i.e., generally, $d(x,y) \geq 0$ if $x \neq y$). Thus, the natural distance matrix can possess entries equal to zero in off-diagonal sites [34].

In the present work, we define for the (shortest path) harmonic centrality measure and the natural harmonic centrality measure their *extended versions*. In our opinion, these derivative centralities should be more effective in identifying influential nodes in complex networks. Namely, for a complex network $G = (V(G), E(G))$, we propose the *extended* (*shortest path*) *harmonic centrality measure*, denoted by EHC, which is given by the following condition

$$EHC(v_i) = \sum_{v_j \in N_G(v_i)} HC(v_j).$$

Analogously, for a complex network $G = (V(G), E(G))$, the *extended natural harmonic centrality measure*, denoted by $ENHC$, is identified with the subsequent expression

$$ENHC(v_i) = \sum_{v_j \in N_G(v_i)} NHC(v_j).$$

Note that both newly defined centrality indices utilize more information than their non-extended counterparts. The extended (shortest path) harmonic centrality measure as well as the extended natural harmonic centrality measure consider not only a focal node but also its nearest neighbors. Therefore, it can be hypothesized that the EHC and $ENHC$ measures will be more effective in determining important vertices in complex networks.

In order to illustrate how the newly defined indices work, consider the small exemplary network G in Figure 1.

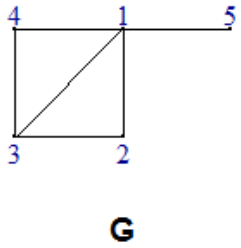

Fig. 1. The sample network G with five nodes

Its natural distance matrix has the form

$$ND(G) = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & 2 & 1.7321 & 2 & 2.2361 \\ 2 & 2 & 0 & 1.7321 & 0 & 1 \\ 3 & 1.7321 & 1.7321 & 0 & 1.7321 & 1.4142 \\ 4 & 2 & 0 & 1.7321 & 0 & 1 \\ 5 & 2.2361 & 1 & 1.4142 & 1 & 0 \end{bmatrix}$$

Then, the nodes of G have the following values of the natural harmonic centrality measure: $NHC(v_1) = 2.0246$, $NHC(v_2) = NHC(v_4) = 2.0774$, $NHC(v_3) = 2.4392$ and $NHC(v_5) = 3.1543$. In turn, the nodes of G possess the subsequent values of the extended natural harmonic centrality measure

$$ENHC(v_1) = NHC(v_2) + NHC(v_3) + NHC(v_4) + NHC(v_5) = 9.7483,$$

$$ENHC(v_2) = ENHC(v_4) = NHC(v_1) + NHC(v_3) = 4.4638,$$

$$ENHC(v_3) = NHC(v_1) + NHC(v_2) + NHC(v_4) = 6.1794,$$

$$ENHC(v_5) = NHC(v_1) = 2.0246.$$

Note that the vertices v_2 and v_4 in the graph G are pseudosimilar. Consequently, $nd_{2,4} = nd_{4,2} = 0$. The values of the extended (shortest path) harmonic centrality measure are calculated analogously.

4. The computational methods and datasets

In order to evaluate the *resolution* (i.e., the *granularity*) of a ranking method, the *monotonicity* ($\mathcal{M}$) and the *percentage of uniquely defined* metric ($\mathcal{PUD}$) of a ranking vector R are used. The monotonicity of R is calculated according to the following formula [15, 16]:

$$\mathcal{M}(R) = \left(1 - \frac{\sum\limits_{r_a \in R} n_{r_a}(n_{r_a} - 1)}{n(n-1)}\right)^2,$$

where n is the size of the ranking vector R and n_{r_a} is the number of ties with the same rank r_a. Note that $\mathcal{M}(R) \in [0, 1]$ for any ranking vector R. This metric quantitatively assesses the fraction of ties in a ranking list. The monotonicity $\mathcal{M}(R)$ is equal to one if the ranking vector R is perfectly monotonic and is equal to zero if all entries in R have the same rank. In turn, the $\mathcal{PUD}$ metric quantifies the fraction of uniquely classified elements in a ranking vector R. The $\mathcal{PUD}$ metric nears 100 % if almost all elements in R have a unique rank and approaches 0 % if almost all elements in R occur in ties with the same rank. Thus, the monotonicity and the $\mathcal{PUD}$ measure allow to evaluate the discrimination ability of rankings induced by centrality indices. However, it should be emphasized that a ranking list with no ties is not necessarily accurate.

It is possible to distinguish two kinds of methods which are employed in order to quantitatively assess the accuracy of a new ranking method. One is based on transmission dynamics and other is based on the network connectivity and the theory that the network damage caused by removing a vertex is equivalent to its importance. Thus, the more significant the node is, the greater influence the node failure triggers [2, 14, 18]. In the present paper, to evaluate the accuracy of the newly proposed ranking methods, the later approach is used. Accordingly, in order to compare the performance of the newly suggested centrality measures with the degree centrality and other state-of-the-art indices, we will scrutinize the importance of nodes for the network connectivity. We will carry out the *deliberate* (i.e., *intensional*, *targeted*) *attack simulations* on the networks and evaluate the effectiveness of the attack strategies by measuring the connectivity and the performance of the post-attack networks. Recall that a deliberate attack means that the network nodes are selectively deleted according to their importance in descending order of some ranking list. The attack scenarios presented in the present work are based on the newly proposed centrality measures and their alternatives.

To measure the connectivity and the performance of the post-attack networks, we will use three performance metrics that can quantify both the topological and functional characteristics of the networks. Namely, in all post-attack networks, we will quantify the number of connected components, the relative size of the giant component as well as the decline rate of the network efficiency. Removing a node from a network can partition the network into several components which are disconnected from each other. We will denote by $C_k(G)$ the number of connected components in the network G after deleting the top$-k$ most important vertices. Thus, in the intensional attack simulations, we will sequentially remove the nodes from the network according to the importance ranking lists induced by the centrality measures under study and calculate the values of $C_k(G)$ after each step of the attack [15, 16, 42]. Undoubtedly, the higher the value of $C_k(G)$ is, the better the ranking method is. Consequently, this evaluation criterion is based on the assumption that if the deletion of influential vertices selected by a certain centrality index leads to a larger fragmentation of the post-attack network, then the attack strategy based on this centrality will achieve higher $C_k(G)$ values. The second metric evaluating the connectivity of the post-attack networks is identified with the relative size of the giant component [1, 2, 14, 15, 18, 42]. In the present work, it is assumed that the initial networks are connected, thus the initial size of the giant component is n (cf. Table 1). The relative size of the giant component, denoted by $r(k)$, after the removal of the top$-k$ significant nodes is expressed by the following formula

$$r(k) = \frac{\sigma_k}{n},$$

where σ_k is the size (i.e., the number of vertices) of the giant component of the post-attack network. In our simulations, we will sequentially delete the nodes from the network according to the importance ranking lists induced by the centrality measures under study and calculate the values of $r(k)$ after each step of the attack. Undoubtedly, the lower the value of $r(k)$ is, the better the ranking method is. Accordingly, this evaluation criterion hinges on the assumption that if the removal of important vertices singled out by a certain centrality measure brings about a larger decomposition of the post-attack network, then the attack scheme based on this centrality index will achieve lower $r(k)$ results. The third metric used to quantify the connectivity of the post-attack networks is the network efficiency (ε) [11,15,16]. Recall that the efficiency of a complex network reflects its connectivity. The better the network connectivity is, the higher the network efficiency is [7, 8]. For a complex network $G = (V(G), E(G))$, this topological quantity is defined

by the following condition

$$\varepsilon = \frac{1}{n\,(n-1)} \sum_{v_i \neq v_j \in V(G)} \frac{1}{d_{ij}}.$$

The decline rate of the network efficiency after k steps of the attack, denoted by $\varepsilon\,(k)$, is given by the following formula

$$\varepsilon\,(k) = 1 - \frac{\varepsilon_k}{\varepsilon_0},$$

where ε_k is the efficiency of the post-attack network (after k steps of the attack) and ε_0 is the efficiency of the initial network. In our simulations, we will sequentially delete the nodes from the network according to the importance ranking lists induced by the centrality measures under study and calculate the values of $\varepsilon\,(k)$ after each step of the attack. Undoubtedly, the higher the values of $\varepsilon\,(k)$ is, the better the ranking method is. Consequently, this evaluation criterion presupposes that if the removal of influential nodes indicated by a certain centrality index induces a faster decline in the efficiency of the post-attack network, then the attack scenario based on this centrality will achieve higher $\varepsilon\,(k)$ results.

In our intensional attack simulation experiments, for a comprehensive comparison, the parameters $C_k\,(G)$, $r\,(k)$ and $\varepsilon\,(k)$ are summed up from $k = 1$ to k corresponding to about 20 % of the nodes in each studied network [1]. Thus, the quantities $\sum_{k=1}^{k} C_k\,(G)$, $\sum_{k=1}^{k} r\,(k)$ and $\sum_{k=1}^{k} \varepsilon\,(k)$ represent the accumulation of the network damages triggered by the attack strategies after k steps.

In summary, the methodological approach taken in the present work hinges on the assumption that centrality measures which are more efficient in identifying significant nodes in complex networks give rise to more destructive attack scenarios [18]. In turn, the destructiveness of an attack scenario is quantified by the cumulative values of $C_k\,(G)$, $r\,(k)$ and $\varepsilon\,(k)$.

Generally speaking, the Kendall's *tau* (τ) correlation coefficient is used to collate the performance of different topology-based measures. This metric evaluates the ranking consistency of two lists that rank the same set of entities [15, 16]. Suppose that $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are two ranked lists that contain n elements, respectively. Any pair of ranks (x_i, y_i) and (x_j, y_j) is considered to be concordant if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$. In turn, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$ the pair is considered to be discordant. In the case of $x_i = x_j$ and $y_i = y_j$ (i.e., tied pair), the pair is said to be neither concordant nor discordant. In this situation, two rank lists x and y

are considered to be independent. The Kendall's *tau* (τ) correlation coefficient is expressed by the following formula

$$\tau(x,y) = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

where $n_0 = \frac{n(n-1)}{2}$, $n_1 = \frac{\sum_i t_i(t_i-1)}{2}$, $n_2 = \frac{\sum_j t_j(t_j-1)}{2}$ and n_c, n_d refer (respectively) to the number of concordant pairs and the number of discordant pairs, t_i and t_j denote the number of tied values in the i-th and j-th group of ties, respectively. The Kendall's *tau* (τ) correlation coefficient ranges between -1 and 1. In the present work, we consider the following ranges in order to quantitatively assess the strength of the relationship between two rankings [15, 16]:

a) no correlation corresponds to: $\tau = 0$,

b) a low positive (negative) correlation corresponds to: $\tau \in (0, 0.5)$ ($\tau \in (-0.5, 0)$),

c) a moderate positive (negative) correlation corresponds to: $\tau \in [0.5, 0.7)$ ($\tau \in (-0.7, -0.5]$),

d) a high positive (negative) correlation corresponds to: $\tau \in [0.7, 0.9)$ ($\tau \in (-0.9, -0.7]$),

e) a very high positive (negative) correlation corresponds to : $\tau \in [0.9, 1)$ ($\tau \in (-1, -0.9]$),

f) a perfect positive (negative) correlation corresponds to: $\tau = 1$ ($\tau = -1$).

In our study, we use four genome-scale metabolic networks corresponding to four species of bacteria. These networks are denoted by ent, sty, sfl and kpn and correspond to *Enterobacter sp.* 638, *Salmonella enterica subsp. enterica serovar Typhi* CT18, *Shigella flexneri* 301 (serotype 2a) and *Klebsiella pneumoniae subsp. pneumoniae* MGH 78578 (serotype K52), respectively. All networks were retrieved from KEGG (*Kyoto Encyclopedia of Genes and Genomes*) database and are treated as undirected [22]. Their basic statistical properties are presented in Table 1.

In Table 1, the symbols $\langle k \rangle$, $k_{\max}$, $L(\lambda_2)$, $\langle l \rangle$, $\langle C \rangle$, C, A_d, $Q(w)$ refer to the average degree, the maximum degree, the algebraic connectivity, the average path length, the average clustering coefficient, the global clustering coefficient

Table 1
The statistical parameters of four genome-scale metabolic networks

Index	ent	sty	sfl	kpn
n	840	857	818	959
m	1299	1306	1251	1466
$\langle k \rangle$	3.0929	3.0478	3.0587	3.0574
$k_{\max}$	37	39	35	43
$L\,(\lambda_2)$	0.0096	0.0091	0.0096	0.013
$\langle l \rangle$	7.4691	7.6634	7.623	7.5466
$\langle C \rangle$	0.2121	0.2124	0.2186	0.1977
C	0.1668	0.1672	0.1717	0.1553
A_d	0.1087	0.1112	0.126	0.0986
$Q\,(w)$	0.7176	0.7111	0.7147	0.6952

(i.e., the ratio of triangles and connected triples), the degree assortativity and the modularity with respect to the walktrap community finding algorithm.

Four degree-preserving random graph models were generated according to the algorithm proposed by F. Viger and M. Latapy and implemented in [9]. These random networks are denoted by ent.rand, sty.rand, sfl.rand and kpn.rand and correspond to the empirical networks ent, sty, sfl and kpn, respectively. These models possess the degree sequences identical with their real-world counterparts. The above-mentioned algorithm always generates undirected, connected simple networks. The procedure relies on first creating an initial (possibly unconnected) simple undirected network with the prescribed degree sequence. Then some rewiring is carried out to make the network connected. Finally, a Monte-Carlo algorithm is employed to randomize the network. The basic statistical properties of the model networks are listed in Table 2.

Besides the real networks and their random analogues, we also compare the performance of the proposed indices with other ranking methods on two synthetic *Barabási-Albert models* (BA). Each synthetic network has 1000 vertices. These model networks were generated according to the original procedure [9, 10, 27]. Namely, the algorithm starts from a connected network with m_0 nodes and, at each step, a new vertex is added to the network and connected to m existing nodes according to the preferential attachment mechanism. In the present paper, we set the values of m to 2 and 3. Accordingly, we have obtained two Barabási-Albert networks, denoted by $BA\,(m = 2)$ and $BA\,(m = 3)$, respectively [9].

Table 2
The statistical parameters of four random models with the preserved degree distribution. The results are averages based on 100 simulation trials

Index	ent.rand	sty.rand	sfl.rand	kpn.rand
$L\,(\lambda_2)$	0.0624	0.0639	0.0672	0.0645
$\langle l \rangle$	5.3279	5.3955	5.3938	5.4111
$\langle C \rangle$	0.0105	0.0095	0.0096	0.009
C	0.0095	0.0087	0.0088	0.0081
A_d	-0.0275	-0.0273	-0.0309	-0.0305
$Q\,(w)$	0.5046	0.5119	0.5094	0.5081

All computations included in the present paper were conducted in the R programming language [3, 9, 21, 35, 43, 44, 48].

5. The results and discussion

In this Section, we evaluate the extended (shortest path) harmonic centrality measure and the extended natural harmonic centrality measure with respect to the resolution, the accuracy in identifying significant nodes in the synthetic as well as in the real-world networks and their rank correlations with the alternative centrality measures.

5.1. The resolution of the tested ranking methods

Tables 3 and 4 contain the results concerning the resolution of the proposed ranking methods in the Barabási-Albert models for $m = 2$ and $m = 3$.

The granularity is quantified by the monotonicity metric (Table 3) and the $\mathcal{PUD}$ metric (Table 4). From the scores in Tables 3 and 4, it can be observed that, in both model networks, the resolution of the HC measure is the lowest. In turn, the granularity of the NHC, EHC and $ENHC$ measures in both model networks is considerably higher. In the synthetic networks, these three centrality measures produce the rankings with the monotonicity scores above 0.999 and with the $\mathcal{PUD}$ scores above 98.5 %. The NHC and $ENHC$ measures defined on the BA $(m = 3)$ network give rise to the perfectly monotonic importance lists.

Table 3
The resolution of four centrality indices evaluated on two BA models quantified by the monotonicity metric

Index	$BA\,(m=2)$	$BA\,(m=3)$
HC	0.9998	0.9996
NHC	> 0.9999	1
EHC	> 0.9999	> 0.9999
$ENHC$	> 0.9999	1

Table 4
The resolution of four centrality indices evaluated on two BA models quantified by the percentage of uniquely defined metric

Index	$BA\,(m=2)$	$BA\,(m=3)$
HC	89.8	82.1
NHC	99	100
EHC	98.5	98.8
$ENHC$	99.2	100

Tables 5 and 6 include the results concerning the resolution of the newly proposed ranking methods (EHC and $ENHC$) as well as the rankings produced by the DC, SL, I and VKC measures defined on four genome-scale metabolic networks.

The granularity is measured by the monotonicity metric (Table 5) and the $\mathcal{PUD}$ metric (Table 6). From Table 5, it can be seen that, in all empirical networks, the DC measure produces the rankings with the monotonicity scores below 0.6. In turn, the EHC, $ENHC$, SL and I measures give rise to the rankings with the monotonicity scores above 0.99. The VKC measure produces the rankings with the monotonicity above 0.98. Thus, the rankings produced by the DC and VKC measures possess the highest number of ties. With respect to the monotonicity metric, these two centrality indices perform the worst. From Table 6, it can be noticed that, in all metabolic networks, the DC measure uniquely identifies below 1 % of the nodes. In turn, in all real-world graphs, the best results are obtained by the I and EHC centralities. These two indices uniquely classify above 86 % of

Table 5
The resolution of six centrality indices evaluated on four metabolic networks quantified by the monotonicity metric

Index	ent	sty	sfl	kpn
$ENHC$	0.9988	0.9987	0.9986	0.9989
EHC	0.9994	0.9994	0.9994	0.9995
DC	0.5984	0.5907	0.5983	0.5981
SL	0.9947	0.9946	0.9944	0.9948
I	0.9993	0.9993	0.9993	0.9994
VKC	0.9851	0.9842	0.9838	0.9825

Table 6
The resolution of six centrality indices evaluated on four metabolic networks quantified by the percentage of uniquely defined metric

Index	ent	sty	sfl	kpn
$ENHC$	82.381	81.3302	81.6626	83.4202
EHC	87.7381	86.6978	87.0416	88.634
DC	0.5952	0.4667	0.7335	0.9385
SL	39.1667	37.5729	38.8753	40.4588
I	87.1429	86.3477	86.4303	88.1126
VKC	52.381	50.0583	50.8557	50.8863

the nodes. The $ENHC$ measure uniquely describes from 81.3302 % to 83.4202 % of the vertices. Thus, with respect to the $\mathcal{PUD}$ metric, the newly suggested $ENHC$ measure performs slightly worse than the centrality based on the inverse square law. On the other hand, the SL measure uniquely identifies from 37.5729 % to 40.4588 % of the nodes whereas the VKC measure uniquely describes from 50.0583 % to 52.381 % of the vertices. Thus, the rankings induced by the DC and SL measures possess the least fraction of the uniquely defined elements. With respect to the $\mathcal{PUD}$ metric, these two centrality indices perform the worst. Note that with respect to both resolution metrics (i.e., $\mathcal{M}$ and $\mathcal{PUD}$), the granularity achieved by two newly introduced centrality measures in the synthetic networks is considerably higher than the resolution in the empirical networks. Note also that, with respect to both resolution metrics, the DC measure performs the worst.

5.2. The intensional attack simulation experiments in the Barabási-Albert models

Our next step in assessing the performance of the newly proposed ranking methods in identifying important nodes is to determine the impact of deleting the top most significant vertices on the network structure. In our deliberate attack simulation experiments, the top twenty percent of the nodes are deleted one by one from the model and empirical networks according to the importance ranking lists produced by the tested centrality measures. Thus, in each step of the attack, only one node is removed according to the importance ranking list and the quantities $C_k(G)$ (i.e., the number of connected components in the post-attack network), $r(k)$ (i.e., the relative size of the giant component in the post-attack network) and $\varepsilon(k)$ (i.e., the decline rate of the network efficiency in the post-attack networks) are measured. After k steps of the attack, we scrutinize the relation between the values of the parameters $C_k(G)$, $r(k)$ and $\varepsilon(k)$ and the significance of the nodes determined by the tested ranking methods.

In our first series of the targeted attack simulation experiments, we juxtapose two newly proposed ranking methods, i.e., the extended (shortest path) harmonic centrality measure and the extended natural harmonic centrality measure with their non-extended counterparts, i.e., the (shortest path) harmonic measure (HC) and the natural harmonic centrality measure (NHC). These four centrality indices are evaluated on the Barabási-Albert model networks for $m = 2$ and $m = 3$. Table 7 shows the cumulative values of the parameter $C_k(G)$ for four centralities defined on the $BA(m = 2, 3)$ networks.

From this table, it can be observed that, in both model networks, the attack strategies based on the extended versions of the harmonic-type centrality measures are significantly more harmful than the attack scenarios based on the non-extended versions of the harmonic-type centrality measures. Thus, it can be asserted that the $ENHC$ and EHC indices remarkably outperform the NHC and HC measures, respectively. It can also be noted that, among the non-extended versions of the harmonic-type measures, the HC index overwhelmingly outperforms the NHC measure whereas among the extended versions of these metrics, the $ENHC$ index surpasses the EHC measure. The above regularity is valid for both model networks. Thus, from the above presented fragmentation results, it can be deduced that the ranking lists based on the extended harmonic-type centralities are more accurate in identifying essential nodes than their analogues based on the non-extended versions of these measures.

Table 7
The accumulation of the damages in two model networks (quantified by the number of connected components in the post-attack graphs) triggered by the four attack strategies guided by the HC, NHC, EHC and ENHC indices. The most aggressive scenarios are marked in bold

Index	$BA(m=2)$	$BA(m=3)$
	$\sum_{k=1}^{k=200} C_k(G)$	$\sum_{k=1}^{k=200} C_k(G)$
HC	13172	5176
NHC	200	200
EHC	29589	8544
$ENHC$	**31601**	**9365**

After we study the structural damages (quantified by the number of connected components in the post-attack networks) triggered by the removal of the top most important vertices, we move on to scrutinize the impact of deleting the most significant nodes on the relative size of the giant component in the $BA(m=2,3)$ networks. Figure 2 (the upper panel) shows the relationship between the number of nodes removed from the networks and the relative size of the giant components ($r(k)$). Table 8 contains the cumulative values of the parameter $r(k)$ for k from 1 to 200.

Recall that the lower the value $r(k)$ is, the more destructive the attack strategy is and (consequently) the more important the removed node is. From these data, it can be noticed that, in both model networks, the attack scenarios based on the extended versions of the harmonic-type centrality indices are considerably more harmful than the attack strategies based on the non-extended versions of these measures. Hence, it can be concluded that the rankings based on the $ENHC$ and EHC measures prevail over the rankings based on the NHC and HC measures, respectively. It can also be observed that, among the non-extended harmonic-type centralities, the HC measure surpasses the NHC index whereas among the extended versions of these measures, the $ENHC$ metric surpasses the EHC index.

In this place, it should be emphasized that two model networks differ in their vulnerability. Namely, the number of necessary vertices which must be removed in order to destroy the giant component varies from one model to another. In the case of the $BA(m=2)$ network, the removal of the 200 most important nodes

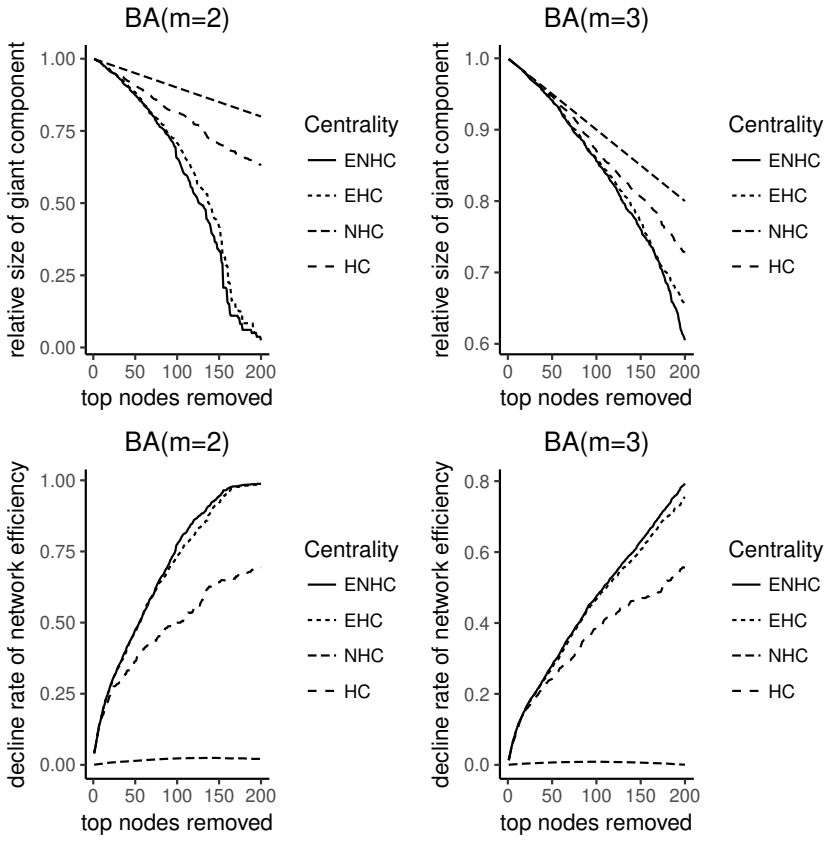

Fig. 2. The relation between the relative size of the giant component and the number of nodes removed from two model networks (the upper panel) as well as between the decline rate of the network efficiency and the number of nodes removed from two model networks (the lower panel)

identified by the $ENHC$ measure brings about high damages to the giant component and thus the decrease of its size by 97.5 %. In turn, in the case of the $BA\,(m = 3)$ network, the deletion of the 200 top nodes indicated by the $ENHC$ measure triggers the decrease of the size of the giant component by 39.5 % (cf. Figure 2, the upper panel). Thus, with respect to the attack strategies based on the $ENHC$ measure, the $BA\,(m = 2)$ model is significantly more vulnerable than the $BA\,(m = 3)$ model.

Figure 2 (the lower panel) presents the relationship between the number of nodes deleted from the networks and the decline rate of the network efficiency. Table 9 lists the cumulative values of the parameter $\varepsilon\,(k)$ for k from 1 to 200.

Table 8
The accumulation of the damages in two model networks (quantified by the relative size of the giant component in the post-attack graphs) triggered by the four attack strategies guided by the HC, NHC, EHC and ENHC indices. The most aggressive scenarios are marked in bold

Index	$BA\,(m=2)$	$BA\,(m=3)$
	$\sum_{k=1}^{k=200} r\,(k)$	$\sum_{k=1}^{k=200} r\,(k)$
HC	161.863	174.436
NHC	179.9	179.9
EHC	122.872	169.889
$ENHC$	**116.859**	**168.484**

Table 9
The accumulation of the damages in two model networks (quantified by the decline rate of the network efficiency in the post-attack graphs) triggered by the four attack strategies guided by the HC, NHC, EHC and ENHC indices. The most aggressive scenarios are marked in bold

Index	$BA\,(m=2)$	$BA\,(m=3)$
	$\sum_{k=1}^{k=200} \varepsilon\,(k)$	$\sum_{k=1}^{k=200} \varepsilon\,(k)$
HC	96.529	70.7205
NHC	3.681	1.1535
EHC	135.5605	88.5091
$ENHC$	**138.395**	**91.1669**

Recall that the higher the value $\varepsilon\,(k)$ is, the more deleterious the attack strategy is and (consequently) the more important the removed node is. From these data, it can be noted that the attack schemes based on the extended harmonic-type centralities are more destructive than the attack strategies based on the non-

extended versions of these indices. Thus, the rankings based on the $ENHC$ and EHC measures are more accurate than the rankings based on the NHC and HC measures, respectively. In turn, among the non-extended harmonic-type centralities, the HC measure overwhelmingly outperforms the NHC measure whereas among the extended harmonic-type centralities, the $ENHC$ index surpasses the EHC measure.

In summary, it can be stated that, with respect to all three evaluation parameters (i.e., $C_k(G)$, $r(k)$ and $\varepsilon(k)$), the extended (shortest path) harmonic centrality measure and the extended natural harmonic centrality measure notably outperform their non-extended versions. Consequently, these newly suggested ranking methods are definitely more efficient in identifying crucial nodes in complex networks than their non-extended counterparts.

5.3. The intensional attack simulation experiments in the genome-scale metabolic networks

In the next series of the intensional attack simulation experiments, we juxtapose two newly proposed centrality measures with four benchmark algorithms, i.e., the degree centrality measure (DC), the local centrality measure (SL), the centrality measure based on the inverse square law (I) and the centrality measure based on the k-shell index and the structural holes (VKC). All six centrality indices are tested on four genome-scale metabolic networks (ent, sty, sfl and kpn).

Table 10 lists the cumulative values of the parameter $C_k(G)$ for six centrality metrics defined on four biochemical networks for k from 1 to about 20 % of the nodes in each network.

Based on the scores included in Table 10, it can be stated that, in all real-world networks, the attack schemes based on the $ENHC$ measure are the most harmful. Thus, with respect to the cumulative values of the parameter $C_k(G)$, the $ENHC$ index outperforms all other centralities. In addition, it can also be observed that, in all datasets, the attack scenarios based on the DC measure and the attack scenarios based on the EHC measure have the second and the third best attack effects, respectively. The scores from Table 10 indicate that, in all cases, when the nodes are deleted according to the importance ranking lists produced by the SL index, the networks have the worst fragmentation effect. Thus, in all metabolic networks, the SL measure performs the worst.

In summary, it can be concluded that the $ENHC$ measure is the most effective in identifying essential nodes. Namely, with respect to the cumulative values of

Table 10
The accumulation of the damages in four metabolic networks (quantified by the number of connected components in the post-attack graphs) triggered by the six attack strategies guided by the ENHC, EHC, DC, SL, I and VKC indices. The most aggressive scenarios are marked in bold

Index	ent	sty	sfl	kpn
	$\sum_{k=1}^{k=170} C_k(G)$	$\sum_{k=1}^{k=175} C_k(G)$	$\sum_{k=1}^{k=165} C_k(G)$	$\sum_{k-1}^{k=195} C_k(G)$
$ENHC$	**27195**	**29234**	**25931**	**36602**
EHC	21528	23170	20154	29092
DC	24359	26173	23285	32596
SL	12981	13659	12448	17460
I	20232	21781	18879	27342
VKC	20726	22080	19540	28177

the parameter $C_k(G)$, the improvement attained by the $ENHC$ measure over the DC index ranges from 10.2 % (the sfl network) to 10.94 % (the kpn network).

In our next series of the measurements, we will study the impact of removing the most vital nodes on the relative size of the giant component in four metabolic networks. Figure 3 presents the relationship between the number of nodes removed from the network and the relative size of the giant component. Table 11 lists the cumulative values of the parameter $r(k)$ for k from 1 to about 20 % of the nodes in each network.

The above data indicate that the attack strategies based on the $ENHC$ measure are the most destructive. Consequently, in all metabolic networks, the $ENHC$ index surpasses all other centralities. In turn, the attack schemes based on the DC measure as well as the attack schemes based on the ENH measure possess the second and the third best attack effect, respectively. On the other hand, the SL measure performs the worst. The above regularities are recorded for all empirical datasets. Based on the above results, it can be asserted that, in all cases, the importance ranking lists induced by the $ENHC$ measure are the most accurate in determining influential nodes. Namely, with respect to the cumulative values of the parameter $r(k)$, the betterment achieved by the $ENHC$ measure over the DC index is in the range from 7.76 % (the sfl network) to 10.36 % (the ent network). Note also that all four networks exhibit the similar vulnerability with respect to the tested attack scenarios. Namely, the deletion of the (about)

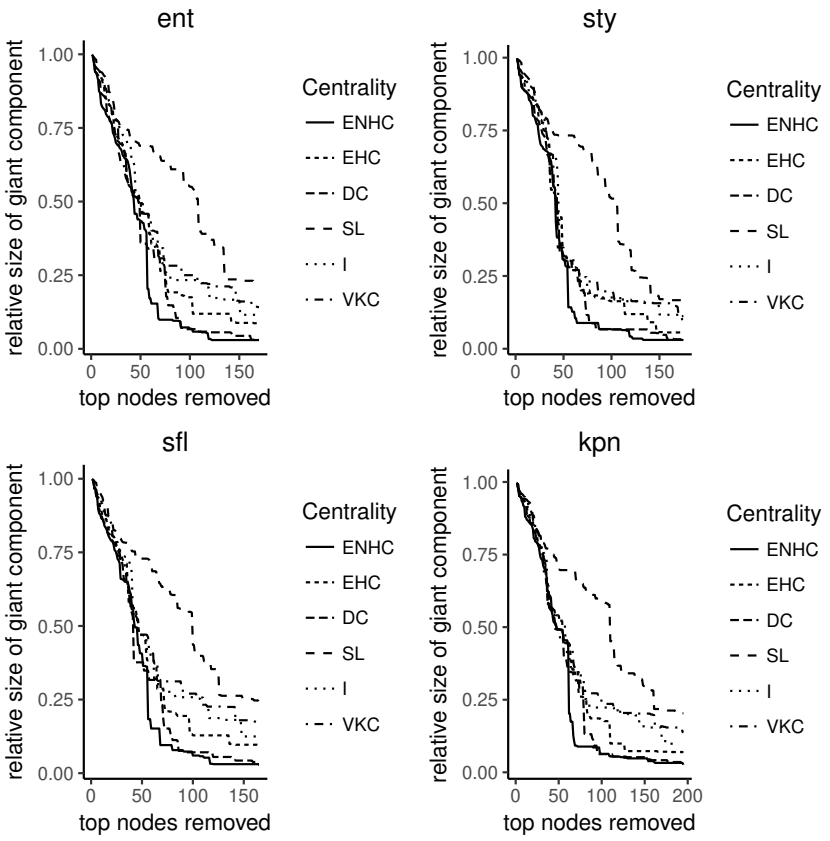

Fig. 3. The relation between the relative size of the giant component and the number of nodes removed from four metabolic networks

20 % of the most important nodes identified by the $ENHC$ measure from the networks ent, sty, sfl and kpn triggers the reduction of their sizes by 97.02 %, 97.08 %, 96.94 % and 97.39 %, respectively. Therefore, with respect to the vulnerability under the attack schemes guided by the $ENHC$ index, the studied metabolic networks are definitely more similar to the Barabási-Albert model for $m = 2$ than for the Barabási-Albert model for $m = 3$.

Figure 4 shows the relationship between the number of nodes removed from the networks and the decline rate of the network efficiency. Table 12 contains the cumulative values of the parameter $\varepsilon(k)$ for k from 1 to about 20 % of the nodes in each network.

Table 11
The accumulation of the damages in four metabolic networks (quantified by the relative size of the giant component in the post-attack graphs) triggered by the six attack strategies guided by the ENHC, EHC, DC, SL, I and VKC indices. The most aggressive scenarios are marked in bold

Index	ent	sty	sfl	kpn
	$\sum_{k=1}^{k=170} r(k)$	$\sum_{k=1}^{k=175} r(k)$	$\sum_{k=1}^{k=165} r(k)$	$\sum_{k-1}^{k=195} r(k)$
$ENHC$	**45.2571**	**43.4936**	**44.7384**	**51.0532**
EHC	58.4083	55.2532	57.9169	62.2711
DC	50.4881	48.1914	48.5024	55.415
SL	95.1107	93.8833	92.9768	104.8446
I	65.9036	60.8086	65.5562	70.6882
VKC	67.3357	61.231	68.0098	74.1366

Table 12
The accumulation of the damages in four metabolic networks (quantified by the decline rate of the network efficiency in the post-attack graphs) triggered by the six attack strategies guided by the ENHC, EHC, DC, SL, I and VKC indices. The most aggressive scenarios are marked in bold

Index	ent	sty	sfl	kpn
	$\sum_{k=1}^{k=170} \varepsilon(k)$	$\sum_{k=1}^{k=175} \varepsilon(k)$	$\sum_{k=1}^{k=165} \varepsilon(k)$	$\sum_{k-1}^{k=195} \varepsilon(k)$
$ENHC$	**146.2508**	**151.5727**	**141.4684**	**169.6323**
EHC	142.3008	147.9262	137.3179	166.3785
DC	144.7085	150.0847	140.2154	168.4407
SL	121.0272	125.7487	117.212	142.2942
I	139.4788	145.5381	134.2337	163.8296
VKC	140.6356	146.2667	134.8871	164.0976

These data demonstrate that, in all datasets, the attack strategies based on the $ENHC$ measure are slightly more deleterious than the attack scenarios based on the DC index. In turn, the attack strategies guided by the DC measure are slightly more harmful than the attack strategies based on the EHC measure. Thus, in all metabolic networks, the $ENHC$ index slightly outperforms the DC measure. On the other hand, the attack strategies guided by the SL measure produce the

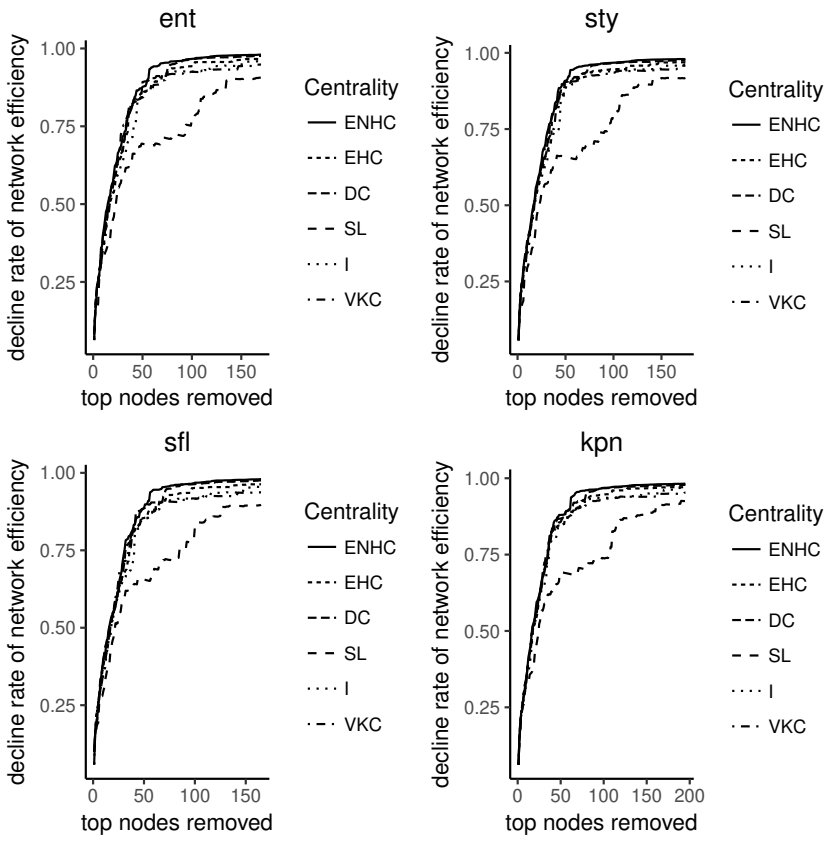

Fig. 4. The relation between the decline rate of the network efficiency and the number of nodes removed from four metabolic networks

poorest results. The above pattern holds in all biochemical networks. Data from Table 12 suggest that, with respect to the cumulative values of the parameter $\varepsilon(k)$, the refinement attained by the $ENHC$ measure over the DC index ranges from 0.7 % (the kpn network) to 1.05 % (the ent network).

To summarize the results of the deliberate attack simulation experiments conducted on four genome-scale metabolic networks, it can be asserted that the attack strategies based on the newly proposed the $ENHC$ measure are the most destructive to the network connectivity. Namely, with respect to three evaluation parameters (i.e., $C_k(G)$, $r(k)$ and $\varepsilon(k)$), the efficacy of the $ENHC$ index outperforms other ranking methods. Thus, the $ENHC$ algorithm possesses the highest identification accuracy. It can also be noted that both newly suggested centrality

indices ($ENHC$, EHC) surpass three representative state-of-the-art centrality measures (SL, I and VKC). Therefore, it can be concluded that the $ENHC$ and EHC algorithms have the ability to effectively distinguish the most essential nodes. Accordingly, the rankings induced by these novel measures are effective in identifying important vertices in the genome-scale metabolic networks and their introduction seems to be justified.

5.4. The rank correlations between the tested ranking methods

The linear and rank correlations between different centrality measures have been studied in many papers ([13, 29, 38, 39] and the references cited therein). It was observed that, in model and empirical networks, the various centrality indices are correlated: influential vertices using one of the measures are frequently also influential using others. For instance, in many cases, vertices with high values of the degree centrality possess also high values of the betweenness centrality [13]. Nonetheless, there are vertices with high values for one centrality and low values for another. In [38], J.R.F. Ronqui and G. Travieso studied the linear correlations quantified by the Pearson correlation coefficient between several centrality metrics and introduced the notion of the *centrality correlation profile* of a given complex network. Namely, the centrality correlation profile of a complex network is identified with a specific pattern of correlations between centrality indices. These authors numerically demonstrated that the centrality correlation profile based on the Pearson correlation coefficient can be used in order to characterize the networks. Namely, this concept allows to distinguish between the empirical networks and their randomly rewired counterparts as well as between the Barabási-Albert models and the Erdős-Rényi models. They concluded that the centrality correlation profile is characteristic property of a given complex network.

Note that, in the present study, we are interested in the orders (i.e., the ranked lists) produced by the centrality measures. Note also that C. Shao and coworkers empirically demonstrated that the rank correlation coefficients (the Kendall's τ or Spearman's ρ) perform better than the Pearson correlation coefficient in scale-free networks [39]. Accordingly, bearing in mind that the metabolic networks are scale-free, we decided to use the Kendall's τau correlation coefficient to quantitatively assess the rank correlations between the studied centrality measures.

Table 13 contains the Kendall's τau correlation coefficients between the non-extended and extended harmonic-type centrality measures evaluated on two Barabási-Albert models for $m = 2$ and $m = 3$.

Table 13
The Kendall rank correlation coefficients between the extended and non-extended harmonic-type centrality indices evaluated on two model networks

Pair of indices	$BA\,(m=2)$	$BA\,(m=3)$
HC/EHC	0.5984	0.631
$NHC/ENHC$	-0.9466	-0.9507

From this table, it can be noticed that, in both model networks, the HC measure is positively moderately correlated with its extended version. In turn, in both model networks, the NHC measure is very strongly negatively correlated with its extended counterpart. Consequently, this very high negative correlation can explain the fact that, in the targeted attack simulation experiments, the $ENHC$ measure overwhelmingly outperforms the NHC index.

Table 14 lists all pairwise Kendall's τau correlation coefficients between six tested centrality measures (i.e., $ENHC$, EHC, DC, SL, I and VKC) defined on four metabolic networks.

All rank correlations contained in Table 14 are positive. From this table, it can be noted that the EHC measure is very highly correlated with the I measure. This high correlation can be easily elucidated by the fact that both indices are based on the shortest path distances. On the other hand, the pairs of measures: $ENHC/DC$, EHC/DC, EHC/VKC, DC/I, SL/I, SL/VKC and I/VKC are highly correlated. Therefore, this implies that vertices which are essential with respect to one definition are, in general, also essential according to other definitions. Two newly proposed indices ($ENHC$ and EHC) are moderately correlated. In turn, the lowest correlation is observed for the pair $ENHC/SL$. Hence, the rankings based on these two methods are divergent. This fact can explain why, in comparison with the attack strategies based on the $ENHC$ index, the attack scenarios guided by the SL metric are the least destructive (cf. Tables 10–12 and Figures 3 and 4). Therefore, we propose to define the *Kendall centrality correlation profile* of a complex network as a specific pattern of rank correlations (quantified by the Kendall's τau correlation coefficient) between centrality measures. From Table 14, it can be observed that all genome-scale metabolic networks are very similar with respect to their Kendall centrality correlation profiles.

In order to check if the Kendall centrality correlation profile of the genome-scale metabolic networks hinges on their degree distribution, we have generated

Table 14
The Kendall rank correlation coefficients between all pairs of six centrality indices (ENHC, EHC, DC, SL, I, VKC) evaluated on four metabolic networks

Pair of indices	ent	sty	sfl	kpn
$ENHC/EHC$	0.5846	0.5746	0.5842	0.5768
$ENHC/DC$	0.8478	0.8498	0.8529	0.8441
$ENHC/SL$	0.2758	0.2609	0.2703	0.2643
$ENHC/I$	0.5096	0.4955	0.5096	0.4990
$ENHC/VKC$	0.4853	0.4664	0.4793	0.4758
EHC/DC	0.8124	0.8027	0.8112	0.8092
EHC/SL	0.6695	0.6631	0.6645	0.6657
EHC/I	0.9078	0.9035	0.9083	0.9047
EHC/VKC	0.7900	0.7714	0.7933	0.7903
DC/SL	0.4981	0.4830	0.4921	0.4862
DC/I	0.7284	0.7137	0.7278	0.7207
DC/VKC	0.6755	0.6547	0.6746	0.6721
SL/I	0.7419	0.7382	0.7366	0.7406
SL/VKC	0.7145	0.7061	0.7196	0.7175
I/VKC	0.8205	0.8012	0.8234	0.8204

for each empirical network the ensemble of 100 model networks which preserve the degree distribution of the original graphs but are less structured. These synthetic networks were constructed *via* the random rewiring procedure (cf. Section 4) and serve as *null models* in our experiments. In comparison with the original metabolic networks, the model networks obtained by the degree-preserving procedure are characterized by the significantly higher algebraic connectivity and the considerably lower values of the average path length, the average clustering coefficient, the global clustering coefficient, the degree assortativity and the modularity with respect to the walktrap partition algorithm (cf. Tables 1 and 2). In order to quantitatively assess the effect of the degree distribution on the Kendall centrality correlation profile in the genome-scale metabolic networks, we measure the Kendall's τau correlation coefficients between six tested centrality measures on the model networks and we calculate the z-score for each pairwise rank correlation. The z-score is calculated according to the following formula

$$z = \frac{\tau_{original} - \langle \tau_{\mathrm{mod}} \rangle}{\sigma\left(\tau_{\mathrm{mod}}\right)},$$

where $\tau_{original}$ is the Kendall's τau correlation coefficient between two centralities measured on the original (i.e., empirical) networks (cf. Table 14), $\langle \tau_{\mathrm{mod}} \rangle$ and $\sigma(\tau_{\mathrm{mod}})$ are the average and the standard deviation of the Kendall's τau correlation coefficients measured on the model networks. Generally, it is assumed that a result is statistically significant if its corresponding z-score is above $|2|$. Thus, in our context, we presuppose that a given rank correlation between two centrality measures quantified by the Kendall's τau correlation coefficient is independent on the degree distribution of the underlying network if its z-score is below -2 or above 2. The Kendall's τau correlation coefficients between all pairs of centrality indices measured on the model networks are contained in Table 15. The corresponding z-scores are listed in Table 16.

Table 15
The Kendall rank correlation coefficients between all pairs of six centrality indices (ENHC, EHC, DC, SL, I, VKC) evaluated on four rewired random network models. The results are averages based on 100 simulation trials

Pair of indices	ent.rand	sty.rand	sfl.rand	kpn.rand
$ENHC/EHC$	0.5930	0.5865	0.5947	0.5915
$ENHC/DC$	0.8583	0.8563	0.8597	0.8562
$ENHC/SL$	0.2521	0.2424	0.2543	0.2397
$ENHC/I$	0.5536	0.5459	0.5564	0.5487
$ENHC/VKC$	0.6364	0.6158	0.6212	0.6208
EHC/DC	0.8646	0.8614	0.8956	0.8634
EHC/SL	0.6296	0.6270	0.6309	0.6198
EHC/I	0.9322	0.9311	0.9346	0.9293
EHC/VKC	0.8104	0.7874	0.7868	0.7992
DC/SL	0.4805	0.4737	0.4831	0.4675
DC/I	0.8168	0.8123	0.8195	0.8116
DC/VKC	0.8055	0.7818	0.7821	0.7903
SL/I	0.6873	0.6855	0.6867	0.6804
SL/VKC	0.5617	0.5617	0.5654	0.5604
I/VKC	0.8019	0.7833	0.7849	0.7934

From these data, it can be seen that, in most cases, the rank correlations between centralities do not depend on the degree distribution of the underlying networks. The highest absolute values of the z-scores are recorded for the rank correlations between the DC measure and two geodesic-based centralities (i.e., EHC and I). Therefore, it can be hypothesized that the centralities based on

Table 16
The z-scores results comparing the rank correlations between six centrality indices (ENHC, EHC, DC, SL, I, VKC) evaluated on the empirical metabolic networks (cf. Table 14) and on the random models with the preserved degree distribution (cf. Table 15)

Pair of indices	ent	sty	sfl	kpn
$ENHC/EHC$	−1.5962	−2.1837	−1.7768	−2.4942
$ENHC/DC$	−3.6440	−2.2295	−2.3252	−3.8238
$ENHC/SL$	1.8120	1.4496	1.1443	1.6934
$ENHC/I$	−5.9597	−7.1971	−6.0147	−6.5763
$ENHC/VKC$	−6.5742	−6.3322	−5.2120	−7.0151
EHC/DC	−16.9604	−19.5085	−15.6695	−18.4598
EHC/SL	3.4094	3.1867	2.6145	3.6049
EHC/I	−8.6569	−9.7598	−8.7213	−7.8646
EHC/VKC	−0.8482	−0.4917	0.1447	−0.3024
DC/SL	1.2284	0.6587	0.5566	1.2109
DC/I	−14.1704	−15.9039	−19.1983	−14.4444
DC/VKC	−4.8452	−3.9091	−2.7238	−4.2105
SL/I	5.2239	5.0899	4.2433	5.3962
SL/VKC	8.4397	8.5060	6.9750	7.5715
I/VKC	1.0710	0.7152	0.9876	1.1436

the topological distance provide unique information which is independent on the degree distribution of the underlying complex networks.

In order to further test if the overall Kendall centrality correlation profile of the metabolic networks depends on their degree sequences, we carry out the Principal Component Analysis (PCA) in which each network (empirical or model) is identified with the 15-dimensional feature vector comprising all pairwise Kendall's τau correlation coefficients. Figure 5 presents four two-dimensional projections of the ensembles consisting of 101 networks (i.e., one real-world network and 100 model networks) from the space defined by the Kendall centrality correlation profile.

In the PCA results, first two principal components explain from 62.9 % (the kpn ensemble) to 66.7 % (the ent ensemble) of the variance contained in the experimental data. In all cases, the empirical networks are located outside the region occupied by their degree-preserving random counterparts. Thus, the PCA outcomes clearly demonstrate that the overall Kendall centrality correlation profile can be used to distinguish between the real-world metabolic networks and their random models having the same degree sequences. Consequently, it can be uttered

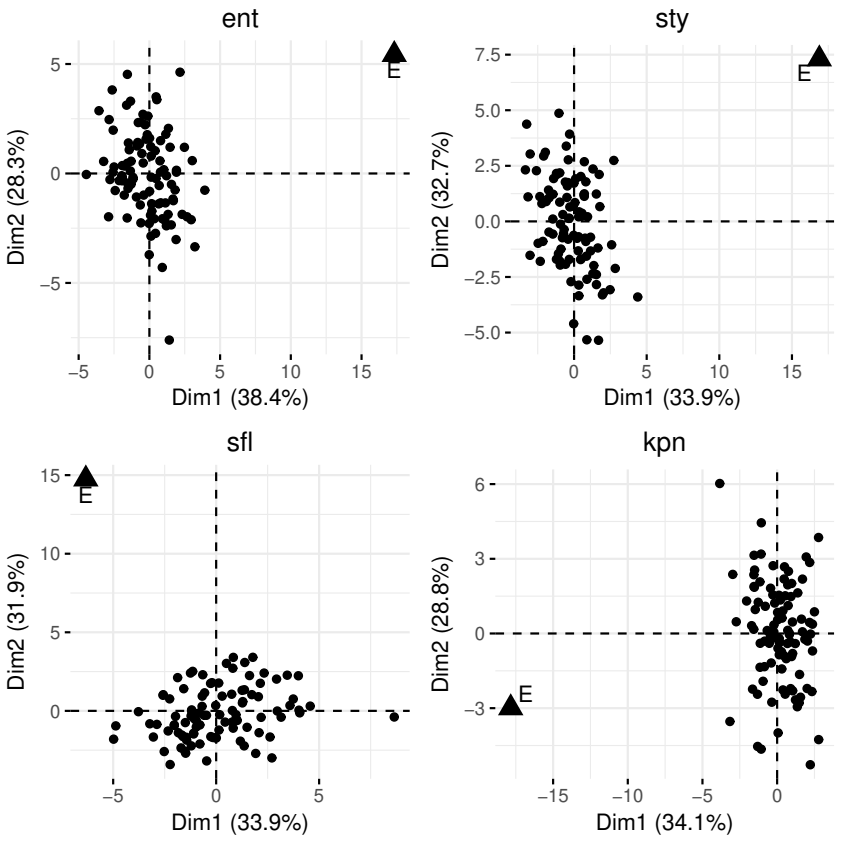

Fig. 5. PCA projections comparing the real-world metabolic networks with their rewired random versions with the preserved degree distribution (in all subfigures, the empirical metabolic network is denoted by the letter E and is marked by the triangle symbol)

that the Kendall centrality correlation profile of the network under study does not depend on their degree distributions and can be employed to characterize these graphs.

Accordingly, the above results show that the concept of the centrality correlation profile of a complex network based on the Pearson correlation coefficients from [38] can be generalized to the rank correlations quantified by the Kendall's τau correlation coefficients. Similarly as the original Pearson-based centrality correlation profile, the Kendall centrality correlation profile can characterize the empirical networks and effectively discriminate them from their random rewirings with the preserved degree sequences.

6. Conclusions

Various centrality indices are commonly used to identify important nodes in complex networks. The different measures correspond to different definitions of the importance of nodes and are applicable to different real-world networks. In the present paper, we have proposed two novel centrality measures – the extended (shortest path) harmonic centrality measure and the extended natural harmonic centrality measure. These metrics can be perceived as refinements of the (non-extended) indices considered in [45]. The numerical results contained in the present work testify that rankings based on both newly suggested centrality measures are characterized by high granularity. The intensional attack simulation experiments carried out on two Barabási-Albert models demonstrated that the newly introduced measures remarkably outperform their non-extended counterparts. In turn, the simulations carried out on four geneome-scale metabolic networks indicated that, in all cases, the attack strategies based on the extended (shortest path) harmonic centrality measure and the extended natural harmonic centrality measure are more destructive than the attack scenarios based on the local centrality measure, the centrality measure based on the inverse square law and the multiattribute centrality measure based on the k-shell index and the structural holes. Thus, in the case of the metabolic networks, both newly introduced centrality indices outperform three representative state-of-the-art centrality metrics. Additionally, the deliberate attack simulation experiments demonstrated that, in the case of the metabolic networks, the attack strategies based on the extended natural harmonic centrality measure are more deleterious than the attack schemes based on the degree centrality. Note that the attack strategies based on the degree centrality are regarded as very aggressive and are commonly used in degrading complex networks [2, 14, 18]. Thus, our results suggest that the extended natural harmonic centrality measure is more effective in dismantling the metabolic networks that the degree centrality. Note also that the problem of the network disintegration is considered in *network science* when we want to disrupt a network if it is harmful [40]. Namely, such phenomena as suppressing the spread of disease, disintegrating the biochemical networks of pathogenic microorganism [4, 28], perturbing cancer networks [32], destabilizing terrorist networks [33], preventing financial contagion [24], controlling the rumor diffusion [41], suppressing the spread of computer viruses are tantamount to the question how to dismantle a given network with a minimal number of node removals. Therefore, it is hoped that the

extended natural harmonic centrality measure will also be effective in dealing with such situations.

The last part of the present report generalizes the notion of the centrality correlation profile of a complex network based on the Pearson correlation coefficient to the case of the rank correlations measured by the Kendall's τau correlation coefficient. The Principal Component Analysis revealed that the redefined centrality correlation profile can be used to distinguish between the empirical metabolic networks and their random models with the preserved degree distribution. Therefore, the Kendall centrality correlation profile can be perceived as a new conceptual framework to characterize complex networks.

In summary, it should be stressed that a centrality measure which is optimal for one complex network is sometimes suboptimal for a different complex network and it is impossible to develop a universal centrality index that best assesses the importance of nodes in every situation. Accordingly, designing efficient algorithms to rank nodes in complex networks with respect to their significance is a long-term challenge.

References

1. Ai X.: *Node importance ranking of complex networks with entropy variation.* Entropy **19** (2017), 303.
2. Albert R., Jeong H., Barabasi A.-L.: *Error and attack tolerance of complex networks.* Nature **406** (2000), 378–382.
3. Auguie B.: *gridExtra: miscellaneous functions for "grids" graphics. R package version 2.3* (2017). https://CRAN.R-project.org/package=gridExtra.
4. Bhattacharya S., Chakrabarti S., Nayak A., Bhattacharya S.K.: *Metabolic networks of microbial systems.* Microb Cell Fact. **2**, no. 3 (2003).
5. Burt R.S.: *Structural holes and good ideas.* Am. J. Soc. **110** (2004), 349–399.
6. Chen D., Lu L., Shang M-S., Zhang Y-C., Zhou T.: *Identifying influential nodes in complex networks.* Physica A **319** (2012), 1777–1787.
7. Crucitti P., Latora V., Marchiori M., Rapisarda A.: *Efficiency of scale-free networks: error and attack tolerance.* Physica A **320** (2003), 622–642.
8. Crucitti P., Latora V., Marchiori M., Rapisarda A.: *Error and attack tolerance of complex networks.* Physica A **340** (2004), 388–394.
9. Csardi G., Nepusz T.: *The igraph software package for complex network research.* InterJournal Complex Syst. **1695** (2006), http://igraph.org.

10. Estrada E.: *The Structure of Complex Networks: Theory and Applications*. Oxford Univ. Press, Oxford 2011.
11. Fei L., Zhang Q., Deng Y.: *Identifying influential nodes in complex networks based on the inverse-square law*. Physica A **512** (2018), 1044–1059.
12. Godsil C.D., Kocay W.L.: *Constructing graphs with pairs of pseudo-similar vertices*. J. Combin. Theory Ser B **32** (1982), 146–155.
13. He X., Meghanathan N.: *Alternative to betweenness centrality: a measure of correlation coefficient*. Proceedings of the Fifth Inernational Conference on Advanced Information Technologies and Application (2016), 1–10.
14. Holme P., Kim B.J., Yoon Ch.N., Han S.K.: *Attack vulnerability of complex networks*. Phys. Rev. E **65** (2002), 056109.
15. Ibnoulouafi A., El Haziti M.: *Density centrality: identifying influential nodes based on area density formula*. Chaos Soliton Fract. **114** (2018), 69–80.
16. Ibnoulouafi A., El Haziti M., Cherifi H.: *M-centrality: identifying key nodes based on global position and local degree variation*. J. Stat. Mech.: Theory and Exp. **2018** (2018), 073407.
17. 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.
18. Iyer S., Killingback T., Sundaram B., Wang Z.: *Attack robustness and centrality of complex networks*. PLoS ONE **8** (4), e59613.
19. Jeong H., Tombor B., Albert R., Oltvai Z.N., Barabási A-L.: *The large-scale organization of metabolic networks*. Nature **407** (2000), 651–654.
20. Jeong H., Manson S.P., Barabasi A.-L., Oltvai Z.N.: *Lethality and centrality in protein networks*. Nature **411** (2001), 41–42.
21. Kassambara A., Mundt F.: *factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.5* (2017). https://CRAN.R-project.org/package=factoextra.
22. KEGG: Kyoto Encyclopedia of Genes and Genomes. www.genome.jp/kegg/.
23. Kitsak M., Gallos L. K., Havlin S., Liljeros F., Muchnik L., Stanley H.E., Makse H.A.: *Identification of influential spreaders in complex networks*. Nature Phys. **6** (2010), 888–893.
24. Kobayashi T., Hasui K.: *Efficient immunization strategies to prevent finantial contagion*. Sci. Rep. **4** (2014), 3834.

25. 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.
26. Koschűtzki D., Junker B.H., Schwender J., Schreiber F.: *Structural analysis of metabolic networks based on flux centrality*. J. Theor. Biol. **265** (2010), 261–269.
27. Latora V., Nicosia V., Russo G.: *Complex Networks. Principles, Methods and Applications*. Cambridge Univ. Press, Cambridge 2017.
28. Lemke N., Herédia F., Barcellos C.K., dos Reis A.N., Mombach J.C.M.: *Essentiality and damage in metabolic networks*. Bioinformatics **20** (2004), 115–119.
29. Li C., Li Q., van Mieghem P., Stanley H. E., Wang H.: *Correlation between centrality metrics and their application to the opinion model*. Eur. Phys. J. B **88** (2015), 65.
30. Marchiori M., Latora V.: *Harmony in the small-world*. Physica A **285** (2000), 539–546.
31. Melak T., Gakkhar S.: *Comparative genome and network centrality analysis to identify drug targets of Mycobacterium tuberculosis H37Rv*, BioMed Res. Int. 2015 (2015), 212061.
32. Quayle A.P., Siddiqui A.S., Jones S.J.M.: *Preferential network perturbation*. Physica A **371** (2006), 823–840.
33. Raab J., Milward H.B.: *Dark networks as problems*. J. Public Adm. Res. Theory **13** (2003), 413–439.
34. Randić M., Pisanski T., Novič M., Plavšić D.: *Novel graph distance matrix*. J. Comput. Chem. **31** (2010), 1832–1841.
35. R Core Team: *R: a language and environment for statistical computing*. R foundation for Statistical Computing, Vienna 2017, http://www.R-project.org/.
36. Ren J., Wang J., Li M., Wu F.: *Discovering essential proteins based on PPI network and protein complex*. Int. J. Data Mining and Bioinformatics **12** (2015), 24–43.
37. Rochat Y.: *Closeness centrality extended to unconnected graphs: the harmonic centrality index*. Application of Social Networks Analysis, Zurich 2009.
38. Ronqui J.R.F., Travieso G.: *Analyzing complex networks through correlations in centrality measurements*. J. Stat. Mech.: Theory Exp. **2015** (2015), P05030

39. Shao C., Cui P., Xun P., Peng Y., Jiang X.: *Rank correlation between centrality metrics in complex networks: an empirical study*. Open Phys. 16 (2018), 1009–1023.

40. Tan S-Y., Wu J., Lu L., Li M-J, Lu X.: *Efficient network disintegration under incomplete information: the comic effect of link prediction*. Sci. Rep. **6** (2016), 22916.

41. Tripathy R.M., Bagchi A., Mehta S.: *A study of rumor control strategies on social networks*. CIKM'10 Proceedings of the 19th ACM Inernational conference on Information and Knowledge Management (2010), 1817–1820.

42. Wenli F., Zhigang L., Ping H.: *Identifying node importance based on information entropy in complex networks*. Phys. Scr. **88** (2013), 065201.

43. Wickham H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York 2009.

44. Wilczek P.: https://github.com/PiotrWilczeknet/R-functions-1.

45. Wilczek P.: *Novel centrality measures and distance-related topological indices in network data mining*. Silesian J. Pure Appl. Math. **7** (2017), 21–63.

46. Wuchty S., Stadler P.F.: *Centers of complex networks*. J. Theor. Biol. **223** (2003), 45–53.

47. Xu H., Zhang J., Yang J., Lun L.: *Identifying important nodes in complex networks based on multiattribute evaluation*. Math. Probl. Eng. 2018 (2018), 8268436.

48. Yang C., Li F.: *RevEcoR: Reverse ecology analysis on microbiome. R package version 0.99.3* (2016). https://CRAN.R-project.org/package=RevEcoR.

49. Zotenko E., Mestre J., O'Leary D.P., Przytycka T.M.: *Why do hubs in the yeast protein interaction network tend to be essential: reaxamining the connection between the network topology and essentiality*. PLoS Comput. Biol. **4**, no. 8 (2008), e1000140.