Journal of Theoretical Biology 263 (2010) 556–565

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Generalized walks-based centrality measures for complex biological networks

Ernesto Estrada

Department of Mathematics and Statistics, Department of Physics, Institute of Complex Systems, University of Strathclyde, Glasgow G1 1XQ, UK

ARTICLE INFO

Article history:
Received 1 October 2009
Received in revised form
3 January 2010
Accepted 14 January 2010
Available online 18 January 2010

Keywords:
Centrality indices
Subgraph centrality
Protein–protein interactions
Complex networks
Matrix functions

ABSTRACT

A strategy for zooming in and out the topological environment of a node in a complex network is developed. This approach is applied here to generalize the subgraph centrality of nodes in complex networks. In this case the zooming in strategy is based on the use of some known matrix functions which allow focusing locally on the environment of a node. When a zooming out strategy is applied new matrix functions are introduced, which give a more global picture of the topological surrounds of a node. These indices permit a modulation of the scales at which the environment of a node influences its centrality. We apply them to the study of 10 protein–protein interaction (PPI) networks. We illustrate the similarities and differences between the generalized subgraph centrality indices as well as among them and some classical centrality measures. We show here that the use of centrality indices based on the zooming in strategy identifies a larger number of essential proteins in the yeast PPI network than any of the other centrality measures studied.

1. Introduction

Complex networks are ubiquitous in biological sciences. They can represent molecular interactions and transformations, such as in transcription networks, protein–protein interaction networks and metabolic networks (Barabási and Oltvai, 2004). Complex networks are also valuable in representing ecological systems, such as in the case of food webs (Jordán and Scheuring, 2004). In both contexts the use of graph theoretic invariants to characterize the local and global topology of these networks is of tremendous importance for extracting useful biological information from such networks (Costa et al., 2007; Jordán et al., 2006, 2007).

Among the graph theoretic invariants characterizing biological networks, centrality indices ranking the relevance of nodes in the network have received great attention in recent years (Costa et al., 2007; Jordán et al., 2007; Junker et al., 2006). In general the notion of node centrality comes from its use in social networks (Chapter 5 in Wasserman and Faust, 1994). Intuitively, it is related to the ability of a node to communicate directly with other nodes, or to its closeness to many other nodes or to the quantity of pairs of nodes which need a specific node as intermediary in their communications. These ideas have materialized in some well-known centrality measures such as degree centrality (DC), closeness centrality (CC), eigenvector centrality (EC) and betweenness centrality (BC) (Wasserman and Faust, 1994). Some of these measures describe the local environment around a node, e.g., degree centrality, and others characterize more globally the position of a node in the network, e.g., eigenvector centrality. An intermediate—neither local nor global—characterization of the node centrality has been claimed as a necessity for the study of biological networks (Jordán and Scheuring, 2002; Jordán et al. 2006). In such "meso-scale" view node centrality should takes into account that the strength of indirect effects decreases with the length of the pathway. In such a way an index accounting for node importance should be between the local and the global scales reflecting far reaching effects but only to smaller and smaller extent (Jordán and Scheuring, 2002; Jordán et al., 2006).

Here we propose a strategy that permits to zooming in and out the topological environment of a node in order to characterize its centrality. This strategy is based on the use of matrix functions which permit to characterize the centrality of a node by taking into account its participation in network's walks. Our strategy permits to give more or less weight to the walks of different lengths producing the desired zooming of the topological environment of the node. Despite the theoretical approaches described here can be applied to different network descriptors we have selected the subgraph centrality (Estrada and Rodríguez-Velázquez, 2005) of a node for illustrating the applications of this local-global focus strategy in complex biological networks. By studying 10 protein–protein interaction (PPI) networks we show that the generalized subgraph centrality indices capture different topological information of the environment of a node. Such information can be useful in making predictions about network-independent functional data of PPI networks. As an example we show here that some of the indices obtained by zooming in the

E-mail address: ernesto.estrada@strath.ac.uk

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.01.014

subgraph centrality are able to identify *in silico* more essential proteins in the yeast PPI network than any other centrality measure.

2. Preliminary definitions

Let $G=(V,E)$ be a network with $|V|=n$ nodes and $|E|=m$ links without loops and multiple edges. A walk of length k is a sequence of (not necessarily different) nodes $v_0, v_1, \ldots, v_{k-1}, v_k$ such that for each $i=1,2,\ldots,k$ there is a link from v_{i-1} to v_i. Let $\mathbf{A}(G)=\mathbf{A}$ be the adjacency matrix of the network. Then, the moment $\mu_k(p,q)=(\mathbf{A}^k)_{pq}$ gives the number of walks of length k starting at the node p and ending at the node q. If $p=q$ the moment $\mu_k(p)$ gives the number of walks starting and ending at the same node, which are known as a self-returning or closed walks (CWs). The total number of CWs in a network is given by the trace of the corresponding power of the adjacency matrix, $\mu_k=tr(\mathbf{A}^k)$. It is known that if $\lambda_1,\lambda_2,\ldots,\lambda_n$ denote the eigenvalues of the adjacency matrix $\mathbf{A}$ of G, then the kth spectral moment of $\mathbf{A}$ can be expressed as follows:

$$\mu_k=tr(\mathbf{A}^k)=\sum_{j=1}^{n}\lambda_j^k. \tag{1}$$

Every CW is related to a given subgraph in a network. For instance, the number of CWs of length 2 for node i equals the degree of node i. CWs of length 3 are related to the number of triangles and CWs of length 4 are related to links, paths of length 2 and squares in the network. Then, the spectral moments of the adjacency matrix are the basis of several structural invariants used for networks in different environments. One of these invariants was introduced to quantify the degree of folding of protein chains (Estrada, 2000) in which a weighted graph is used to represent the adjacency between dihedral angles in the protein backbone. This index was later generalized to any complex network as a way to quantify the content of subgraphs in the network. It is defined by the Taylor expansion of the spectral moments of the form (Estrada, 2000; Estrada and Rodríguez-Velázquez, 2005)

$$EE(G)=\mu_0+\mu_1+\frac{\mu_2}{2!}+\frac{\mu_3}{3!}+\cdots+\frac{\mu_k}{k!}+\cdots, \tag{2}$$

which has the following closed form in terms of the graph spectrum (Estrada, 2000; Estrada and Rodríguez-Velázquez, 2005a)

$$EE(G)=\sum_{j=1}^{n}e^{\lambda_j}=tr(e^{\mathbf{A}}), \tag{3}$$

where the *exponential adjacency matrix* is defined as

$$e^{\mathbf{A}}=\mathbf{I}+\mathbf{A}+\frac{\mathbf{A}^2}{2!}+\frac{\mathbf{A}^3}{3!}+\cdots+\frac{\mathbf{A}^k}{k!}+\cdots. \tag{4}$$

The number of CWs of length 2, which are accounted by $\mathbf{A}^2$, are counted twice for every link in the network. Similarly, the CWs of length 3, which are accounted by $\mathbf{A}^3$, are counted six times for every triangle. This can give the false impression that the penalization scheme used in (4) is based on this fact. That is, that CWs of length 2 are penalized by 2! and CWs of length 3 are penalized by 3!, because these are the number of repetitions of these CWs for links and triangles, respectively. However, this is not true for other powers of the adjacency matrix beyond $\mathbf{A}^3$. For instance, for $\mathbf{A}^4$ every link contributes twice, every path of length 2 contributes four times and every square contributes eight times, which is not equal to 4!. In the next section we are going to explain what are the conditions that the penalization used need to fulfill in order to produce appropriate descriptions of the subgraph centrality.

The index $EE(G)$ was proposed as a subgraph centralization of the network and it is nowadays referred to as the Estrada index of a graph (de la Peña et al., 2007; Carbó-Dorca, 2008; Deng et al., 2009).

The subgraph centrality of the node p is given by $(e^{\mathbf{A}})_{pp}$ and it has the following spectral formula (Estrada and Rodríguez-Velázquez, 2005),

$$EE(p)=\sum_{k=0}^{\infty}\frac{\mu_k(p)}{k!}=(e^{\mathbf{A}})_{pp}=\sum_{j=1}^{n}[\phi_j(p)]^2e^{\lambda_j}, \tag{5}$$

where $\mu_k(p)=(\mathbf{A}^k)_{pp}$ is the number of CWs starting (and ending) at node p, $\phi_j(p)$ is the pth entry of the jth orthonormal eigenvector $\boldsymbol{\varphi}_j$ associated to the eigenvalue λ_j.

Recently Estrada and Higham (2008) proposed a general formulation for the invariants based on Taylor series expansion of spectral moments,

$$EE(G,c)=\sum_{k=0}^{\infty}c_k\mu_k. \tag{6}$$

This general formulation was applied to complex networks by considering the following invariant,

$$EE(G,c)=\sum_{k=0}^{\infty}\frac{1}{(n-1)^k}\mu_k, \tag{7}$$

which eventually converges to the trace of the resolvent of the adjacency matrix (Estrada and Higham, 2008),

$$EE(G,c)=tr\left(\mathbf{I}-\frac{1}{n-1}\mathbf{A}\right)^{-1}=\sum_{j=1}^{n}\left(1-\frac{\lambda_j}{n-1}\right)^{-1}. \tag{8}$$

The coefficients c_k used in this index largely penalize long closed walks in the graph. As a consequence, $EE(G,c)$ is very similar to the degree centrality of a node making it a very local index in comparison with the subgraph centrality. As we have explained before our aim here is to produce a method that permit us to zooming in and out the topological environment of a node in such a way that we can obtain indices in between the local-global characterization of node centrality, i.e., to obtain a sort of "meso-scale" centrality indices.

3. On walk-based network measures

The concept of node centrality in networks was introduced in the beginning of the 50's when Bavelas (1948, 1950) and Leavitt (1951) used this concept in communication networks. Freeman introduced the degree centrality in social networks in 1979 (Freeman, 1979), which can be written in matrix form as

$$k_i=(\mathbf{Au})_i, \tag{9}$$

where $\mathbf{u}$ is a unit column vector.

Instead of using only the adjacency matrix, Katz (1953) proposed to use the different powers of it multiplied by certain coefficients, which can be considered the first walk-based characterization of nodes in a network

$$K_i=\{[(\mathbf{I}-\alpha\mathbf{A})^{-1}-\mathbf{I}]^T\mathbf{u}\}_i, \tag{10}$$

where $\mathbf{I}$ is the identity matrix and $\alpha<1/\lambda_1$.

Then, a quantum leap appears when Bonacich (1972, 1987) introduced the definition of eigenvector centrality in which nodes' centrality is a function of the centrality values of adjacent nodes,

$$\phi_1(i)=\left(\frac{1}{\lambda_1}\mathbf{A}\boldsymbol{\varphi}_1\right)_i, \tag{11}$$

where λ_1 and $\boldsymbol{\varphi}_1$ are the Perron–Frobenius eigenvalue and eigenvector or $\mathbf{A}$, respectively. The reader can find more details about centrality measures in the excellent review of Borgatti and Everett (2006) as well as in the Chapter 5 of the book by Wasserman and Faust (1994).

The walk-based centrality measures can be expressed by the general formula $f(\mathbf{A})\mathbf{u}$, where the matrix function $f(\mathbf{A})$ is given by

$$f(\mathbf{A}) = \sum_{k=0}^{\infty} c_k \mathbf{A}^k. \tag{12}$$

It is straightforward to realize that the degree, eigenvector and subgraph centrality can be considered as walk-based characterizations of the nodes in complex networks. For instance, in this context the Bonacich eigenvector centrality in vector form can be written as,

$$\boldsymbol{\varphi}_1 = \left\{ \lim_{k\to\infty} \left[\frac{1}{k} (\mathbf{A} + \lambda_1^{-1}\mathbf{A}^2 + \lambda_1^{-2}\mathbf{A}^3 + \cdots + \lambda_1^{-k}\mathbf{A}^{k+1}) \right] \right\} \mathbf{u}. \tag{13}$$

The only conditions that the coefficients c_k needs to fulfill can be resumed as follows: (i) the coefficients need to make the series (12) convergent; (ii) the coefficients need to give more weights to shortest walks than to longer ones; (iii) the indices produced need to be real-positive numbers.

It is worth mentioning that several other successful centrality measures are not based on walks, such as the closeness, betweenness, and information centrality (Borgatti and Everett, 2006).

Here we propose to generalize these invariants by using a new strategy that allows the definition of new centrality measures containing structural information on biological networks which is not duplicated by other measures. The duplication of information is analyzed by considering the strength of correlation among these measures, which indicates how much overlap exists among their information content.

4. Strategy for generalization

The motivation for introducing the subgraph centralization and centrality was to count the participation of a node in all the subgraphs existing in the graph. This was accomplished by considering the number of closed walks in which a node takes place giving larger weights to shorter closed walks. For the sake of simplicity let us consider the graph illustrated in Fig. 1. The subgraph centrality clearly identifies the node 5 as the most central one in the graph followed by the node 7. The main difference between these two nodes is given by their participation in paths of length 2 (2-paths) as they take part in the same number of triangles and squares. Node 5 appears six times as the center and six times as an end point of a 2-path. However, node 7 appears three times as a center and five times as an end point of a 2-path. Node 1 is ranked fourth by the subgraph centrality after the equivalent nodes 2 and 4. This node takes part only in one triangle (1-2-4) instead of two (5-8-7, 5-6-7). However, it appears seven times as an end point of a 2-path and three times as its center.

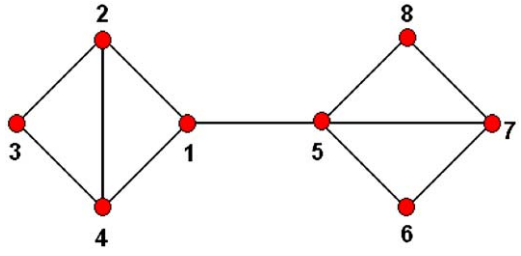

Fig. 1. Graph used to illustrate the necessity for rescaling the subgraph centrality and related indices.

2-paths can be very relevant in clustered networks as the one represented by the graph in Fig. 1. For instance, the two nodes taking part in the largest number of 2-paths are nodes 1 and 5, which here form a bridge between the two clusters in the network. On the other hand, because the number of 1-paths in which a node takes place is the degree of the node, the number of 2-paths can be seen as the next step in extending such centrality. The same reason can be generalized to consider m-paths ($m \geq 2$). A strategy for giving more weights to these fragments is to decrease the penalization imposed to the smaller spectral moments in the original definition of the subgraph centrality. For instance, let us use the following expression to calculate a modified subgraph centrality:

$$\mu_0(i) + \mu_1(i) + \mu_2(i) + \mu_3(i) + \mu_4(i) + \mu_5(i) + \mu_6(i) + \frac{\mu_7(i)}{2!} + \frac{\mu_8(i)}{3!} + \cdots.$$

Then, the new ranking of the nodes places node 1 as the second most central in the graph following node 5 and preceding node 7. The differences in the ranking are more evident if we consider the effect of removing the most central nodes according to the original subgraph centrality and the modified one. Removing node 7 does not affect the connectivity of the graph while removing node 1 produces the same effect as the removal of node 5, which separates the graph into two disjoint components. This example illustrates how by modulating the penalization of the spectral moments we can change the ranking of the nodes in a graph according to the relevance of different structural elements in which they participate.

We can generalize this idea by considering positive and negative rescaling of the expression (5). By positive (negative) rescaling we mean moving forward (backward) the numerators (spectral moments) respect to the denominators (factorials) in the expression (2) as illustrated below.

positive rescaling:

$$\frac{\mu_0}{0!} + \frac{\mu_1}{1!} + \frac{\mu_2}{2!} + \cdots \Rightarrow \frac{\mu_0}{1!} + \frac{\mu_1}{2!} + \frac{\mu_2}{3!} + \cdots.$$

negative rescaling:

$$\frac{\mu_0}{0!} + \frac{\mu_1}{1!} + \frac{\mu_2}{2!} + \cdots \Rightarrow \mu_0 + \frac{\mu_1}{0!} + \frac{\mu_2}{1!} + \cdots.$$

If we consider the CWs accounted for by both kinds of rescaling we can see that the positive rescaling corresponds to zooming in of the environment of a node. That is, by penalizing more the longest walks we concentrate more in the local environment of the corresponding node. On the other hand, the negative rescaling corresponds to a zooming out of the surrounds of the corresponding node. In this case we allow long walks to contribute to the index in such a way that we obtain more global information about the environment of the node under study.

5. Zooming in by positive rescaling

By moving forward one step the spectral moments respect to the factorial denominators we obtain the following Taylor series

$$EE^1(G) = \mu_0 + \frac{\mu_1}{2!} + \frac{\mu_2}{3!} + \frac{\mu_3}{4!} + \cdots + \frac{\mu_k}{(k+1)!} + \cdots. \tag{14}$$

Then the index $EE^1(G)$ has the following spectral formula only in the case when no eigenvalue is equal to zero

$$EE^1(G) = \sum_{j=1}^{n} \frac{e^{\lambda_j} - 1}{\lambda_j}. \tag{15}$$

It is straightforward to realize that $EE^1(G)$ can be obtained as the trace of the $\psi_1(\mathbf{A})$ matrix function (Higham, 2008),

$$EE^1(G) = tr\,\psi_1(\mathbf{A}) = tr\left(\frac{e^{\mathbf{A}}-\mathbf{I}}{\mathbf{A}}\right), \tag{16}$$

where $\mathbf{A}$ is a non-singular matrix. The function $\psi_1(\mathbf{A})$ appears in the exact solution of ordinary differential equations (Hochbruck et al., 1998). For instance, let

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}\mathbf{y}+\mathbf{b} \quad \text{for } t>0,$$

$$\mathbf{y}(0) = \mathbf{y}_0.$$

Then, the function $\psi_1(\mathbf{A})$ appears in the exact solution of this equation when the square matrix $\mathbf{A}$ and the column vector $\mathbf{b}$ are independent of t,

$$y(t) = y_0 + t\psi_1(t\mathbf{A})(b+\mathbf{A}y_0).$$

The importance of this relation is evident by considering that $EE^0(G)$ represents the partition function of a network obtained by solving the Schrödinger equation in which the Hamiltonian is simply the negative adjacency matrix (Estrada and Hatano, 2007). Then, the first order index is related to the solutions of the nonlinear Schrödinger equation using exponential integrators (Hochbruck and Lubich, 1999).

The index $EE^1(G)$ can be obtained by means of the following recurrence relation:

$$EE^1(G) = EE^0(G) - \sum_{k=1}^{\infty}\frac{\mu_k}{(k+1)(k-1)!}, \tag{17}$$

where the second member of the RHS of this expression converges as follow when no eigenvalue is equal to zero

$$\sum_{k=1}^{\infty}\frac{\mu_k}{(k+1)(k-1)!} = \sum_{j=1}^{n}\frac{\lambda_j e^{\lambda_j}-e^{\lambda_j}+1}{\lambda_j}. \tag{18}$$

The application of this power sum to the adjacency matrix of the graph gives rise to a new matrix function, which when $\mathbf{A}$ is non-singular can be expressed as

$$\sum_{k=1}^{\infty}\frac{\mathbf{A}^k}{(k+1)(k-1)!} = \frac{\mathbf{A}e^{\mathbf{A}}-e^{\mathbf{A}}+\mathbf{I}}{\mathbf{A}}.$$

We can extend the positive rescaling approach to generate a series of indices characterizing a graph in terms of the spectral moments of the adjacency matrix weighted by inverse factorials. The general formulation for these indices is given below:

$$EE^t(G) = \sum_{k=0}^{\infty}\frac{\mu_k}{(k+t)!}. \tag{19}$$

The generalized $EE^t(G)$ index has the following spectral formula when no eigenvalue is equal to zero:

$$EE^t(G) = \sum_{j=1}^{n}\frac{e^{\lambda_j}-\sum_{s=1}^{t}\frac{\lambda_j^{t-s}}{(t-s)!}}{(\lambda_j)^t}. \tag{20}$$

These indices are also related to matrix functions through the trace formula:

$$EE^t(G) = tr\psi_t(\mathbf{A}), \tag{21}$$

where the $\psi_t(\mathbf{A})$ matrix functions (Higham, 2008) have the following integral formula:

$$\psi_t(\mathbf{A}) = \frac{1}{(t-1)!}\int_0^1 e^{(1-t)\mathbf{A}}x^{t-1}\,dx. \tag{22}$$

The following recurrence formula is known for these matrix functions (Higham, 2008),

$$\psi_t(\mathbf{A}) = \mathbf{A}\psi_{t+1}(\mathbf{A}) + \frac{1}{t!}.$$

The $EE^{t+1}(G)$ index can be obtained by means of the following recurrence relation:

$$EE^{t+1}(G) = EE^t(G) - \sum_{k=1}^{\infty}\frac{\mu_k}{(k+t+1)(k+t-1)!}, \tag{23}$$

where the Taylor series in second member of the RHS of this expression converges as follow when no eigenvalue is equal to zero,

$$\sum_{k=1}^{\infty}\frac{\mu_k}{(k+t+1)(k+t-1)!} = \sum_{j=1}^{n}\left(\frac{\lambda_j e^{\lambda_j}-e^{\lambda_j}+1}{\lambda_j^{t+1}} - \frac{\sum_{s=2}^{r}\frac{1}{s}\lambda_j^s}{\lambda_j^{t+1}}\right). \tag{24}$$

It should be mentioned that the functions $\psi_t(\mathbf{A})$ are entire and they can always be calculated despite some eigenvalues of the adjacency matrix are equal to zero or the matrices are singular, because they can be represented as a power series which converges compactly. There are several numerical approaches to address the calculation of these functions which have been reported in the literature and the reader is referred to it for details (Higham, 2008).

6. Zooming out by negative rescaling

In the negative rescaling approach we are interested in not penalizing the closed walks of the smallest length in the graph. In the original $EE^0(G)$ the spectral moments of length 0 and 1 are not penalized by dividing them with any factor. Suppose we are interested in extending this idea to the second spectral moments, such that we have the following power sum:

$$\mu_0+\mu_1+\mu_2+\frac{\mu_3}{2!}+\frac{\mu_4}{3!}+\cdots,$$

or to the third ones in such a way that we have

$$\mu_0+\mu_1+\mu_2+\mu_3+\frac{\mu_4}{2!}+\frac{\mu_5}{3!}+\cdots.$$

In general, we can define the following negatively rescaled indices

$$EE^{-t}(G) = \sum_{s=0}^{t-1}\mu_s + \sum_{k=t}^{\infty}\frac{\mu_k}{(k-t)!}, \tag{25}$$

which have the following spectral realization:

$$EE^{-t}(p) = \sum_{j=1}^{n}\left(\sum_{s=0}^{t-1}\lambda_j^s + \lambda_j^t e^{\lambda_j}\right). \tag{26}$$

The $EE^{-t}(G)$ index can be obtained by means of the following recurrence relation:

$$EE^{-t}(G) = EE^{-t+1}(G) + \sum_{k=t+1}^{\infty}\frac{\mu_k}{(k-t+1)(k-t-1)!}, \tag{27}$$

where the Taylor series in the second member of the RHS converges as follow:

$$\sum_{k=t+1}^{\infty}\frac{\mu_k}{(k-t+1)(k-t-1)!} = \sum_{j=1}^{n}\lambda_j^{t-1}(1+\lambda_j e^{\lambda_j}-e^{\lambda_j}). \tag{28}$$

By applying this Taylor series to the adjacency matrix of a graph we can obtain a new matrix function, which is given by

$$\mathbf{A}^t(\mathbf{I}+\mathbf{A}e^{\mathbf{A}}-e^{\mathbf{A}}). \tag{29}$$

7. Study of protein–protein interaction networks

We study here 10 protein–protein interaction (PPI) networks in which nodes represent proteins and links represent interactions between pairs of proteins. These PPI networks correspond to *Archaeglobus fulgidus* (Motz et al., 2002), Kaposi sarcoma-associated herpes virus (KSHV) (Uetz et al., 2006), varicella-zoster virus (VZV) (Uetz et al., 2006), *Bacillus subtilis* (Noirot and Noirot-Gros, 2004; Hoebeke et al., 2001), *Escherichia coli* (Butland et al., 2005), malaria parasite *Plasmodium falciparum* (LaCount et al., 2005), the worm *Caenorhabditis elegans* (Li et al., 2004), *Helicobacter pylori* (Rain et al., 2001), *Saccharomyces cerevisiae* (von Mering et al., 2002; Bu et al., 2003), and *Homo sapiens* (Rual et al., 2005). Here we consider the main connected component of these networks in which interactions between proteins are taken to be undirected and no self-interactions are considered. Consequently, the corresponding networks are undirected and do not contain self-loops.

We study here the subgraph centrality of the protein i in a PPI network, which is given by $EE^0(i)=(e^{\mathbf{A}})_{ii}$. The subgraph centrality has been previously applied to the identification of essential proteins in proteomic maps (Estrada, 2006a, 2006b; Zotenko et al., 2008; Lin et al., 2008; Gursoy et al., 2008) and the characterization of malignant tissues (Platzer et al., 2007). It has also been applied to the study of weighted graphs to account for the degree of folding of protein chains (see for instance Estrada, 2002, 2004) as well as to describe the molecular structure of drug-like and environmentally relevant organic compounds (for a review see Estrada and Uriarte, 2001).

We are interested in comparing this index with the generalized subgraph centrality indices defined by using positive and negative rescaling, which are given by the following expressions, respectively,

$$EE^{t}(i)=[\psi_t(\mathbf{A})]_{ii},$$

$$EE^{-t}(i)=(\sum_{s=0}^{t-1}\mathbf{A}^s+\mathbf{A}^t e^{\mathbf{A}})_{ii},$$

where the matrix functions $\psi_t(\mathbf{A})$ were previously defined.

First we analyze the intercorrelation between all subgraph centrality indices obtained here for $-10\le t\le 10$. It is easy to realize that $EE^{-t}(i)$ diverges as $t\to\infty$. On the other hand, $\lim_{t\to\infty}\psi_t(\mathbf{A})=\mathbf{0}$, where $\mathbf{0}$ is an all-zeros matrix, and consequently $EE^{t}(i)$ tends to zero as $t\to\infty$. We have observed that for the PPI networks analyzed here this already happens for $t\ge 8$. Then, we exclude these values from our analysis. As a non-parametric measure of correlation between the indices we use the *Kendall* τ statistics (Kendall, 1938), which represents the difference between the probability that the observed data are in the same order for the two variables versus the probability that the observed data are in different orders for the two variables. Let p_c and p_d be the number of concordant and discordant pairs of data points, respectively, such that $p=p_c+p_d$. Then the Kendall τ index is defined as (Kendall, 1938)

$$\tau=\frac{2(p_c-p_d)}{p(p-1)}.$$

The values of τ for every pair of index are represented as a correlation matrix C. Then we use the first eigenvalue ε_1 of C normalized by the number of data points p as a measure of the global intercorrelation between the indices studied. Note that $0\le\varepsilon_1/p\le 1$.

In Table 1 we present the values of the intercorrelation measures between the subgraph centrality indices for the PPI networks studied here. As can be seen there are relatively large intercorrelations between the indices defined here, which is expected from the fact that they are measuring the same topological properties of nodes, i.e., their subgraph centrality. Most of this intercorrelation arises from the indices which are obtained using the same strategy, i.e., positive or negative rescaling. We have used factor analysis for studying the subgraph centrality indices for the nodes of the PPI network of *C. elegans*. In Fig. 2A it can be seen that the first two factors divide the walk-based indices into two clusters corresponding to the positively and negatively rescaled indices, respectively. Similar patterns are also observed in the correlation matrices for this (see Fig. 2B) and the other PPI networks studied.

Table 1
Intercorrelation between the generalized subgraph centrality indices EE^t, $-10\le t\le 7$, for 10 PPI networks.

No.	PPI network	N	E	ε_1/p
1	*A. fulgidus*	32	36	0.788
2	KSHV	50	114	0.902
3	VZV	53	148	0.933
4	*B. subtilis*	79	92	0.775
5	*P. falciparum*	229	604	0.875
6	*E. coli*	230	695	0.924
7	*C. elegans*	314	363	0.778
8	*H. pylori*	710	1396	0.871
9	*S. cereviciae*	2224	6609	0.924
10	*H. sapiens*	2783	6007	0.816

The number of proteins (N), the number of interactions (E) and the first eigenvalue of the correlation matrix based on the Kendall τ indices (ε_1) normalized by the number of observations (p) are given.

The second part of this global analysis of the generalized subgraph centrality indices is devoted to compare them with some of the best known centrality indices. These "classical" centrality indices are the degree (DC), closeness (CC), betweenness (BC), and the eigenvector centrality (EC). The reader is referred to the Chapter 5 of the book of Wasserman and Faust (1994) to obtain details about these indices. While the degree and eigenvector centrality are clearly walk-based centrality indices, the closeness and betweenness are based on the concept of shortest path distance.

In Fig. 3 we show the Kendall indices for the non-parametric correlations between the generalized centrality indices and the four classical centrality indices studied. The first interesting observation is that the negatively rescaled subgraph centrality indices are highly correlated with the eigenvector centrality and these correlations decay when the value of t increases. Such strong correlations can be explained by the fact that the negatively rescaled indices do not penalize so heavily long walks. For instance, in the index $EE^{-10}(i)$ there is no penalization for walks of length between 1 and 11. It is known that for non-bipartite connected graphs as the ones studied here, the eigenvector centrality of a node is proportional to the number of walks of length k starting at this node as $k\to\infty$, which explains the empirical correlation obtained. The second interesting observation is that the correlation between the degree centrality and the generalized subgraph centralities increases as the values of t increases. That is, negatively rescaled indices are less correlated with the degree than the positively rescaled ones. This observation is easily explained by the fact that in the positively rescaled indices we penalize very much long walks. Consequently, walks of length 2, which are equal to the degree of the node, have a large influence of these indices explaining the observed empirical correlations. Finally, the closeness centrality displays a similar behavior to the eigenvector centrality and the betweenness centrality is more similar to the degree in their correlations with the generalized subgraph centrality indices.

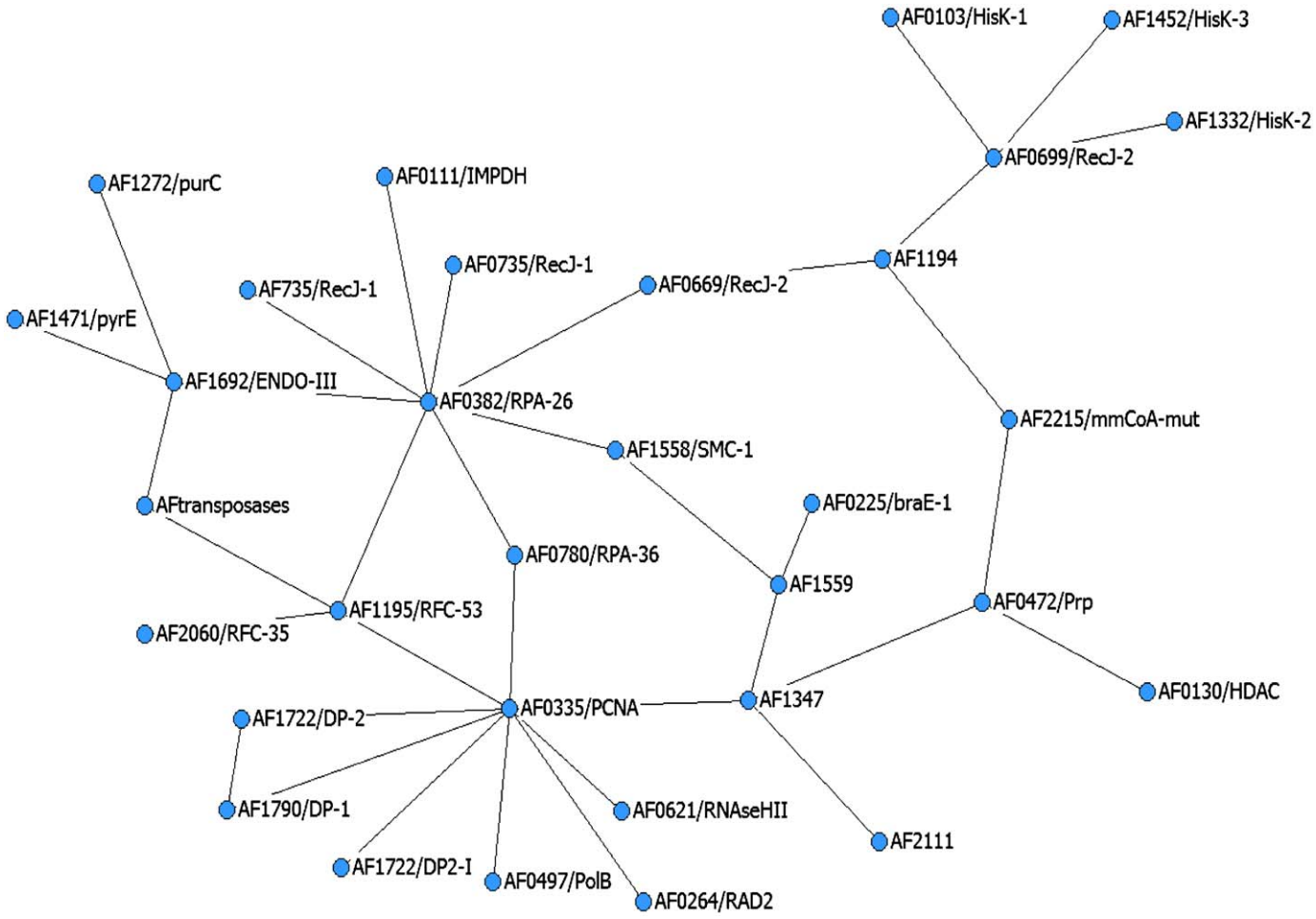

Fig. 2. Protein interaction network of the archaebacterium *A. fulgidus*.

Based on these empirical findings we can conclude that the negatively rescaled subgraph centrality indices represent a zooming out which describe more globally the environment around a node due to their small penalization of the longer walks. On the other hand, the positively rescaled subgraph centrality measures are describing more local characteristics of the surrounds of a node due to the heavy penalization of long walks.

In order to have a closer look to the differences between the walk-based indices introduced here and the subgraph centrality we are going to study the PPI network of *A. fulgidus* in more detail (see Fig. 4 for illustration). The subgraph centrality index $EE^0(i)$ identifies proteins AF0335 and AF0382 as the most central ones in this proteome. These two proteins are also the most connected ones. The first of them is involved in tethering polymerases to DNA and the second one in replication. The ranking is followed by AF1195 (replication factor), AF1347 (unknown function), AF1692 (endonuclease III) and AF0699 (DNA-specific exonuclease).

The indices $EE^t(i)$ are in general strongly correlated to $EE^0(i)$ with Kendall τ indices higher than 0.90. However, a looking glass analysis reveals certain differences. For instance, we can see that the first six proteins ranked by the index $EE^3(i)$ are exactly the same as the ones appearing in the ranking due to $EE^0(i)$. Now, in the ranking produced by $EE^0(i)$ the seventh to tenth proteins are AF1790, AF1722 and AF0780, while in the one produced by $EE^3(i)$ are AF1194, AF1559 and AF0472. These differences are quite significant from the topological point of view. For instance, the deletion of the protein AF1790 ranked as the seventh by $EE^0(i)$ does not produce any other topological change in the protein interaction network. However, removing protein AF1194, which is ranked seventh by $EE^3(i)$ produces the disconnection of four other proteins. Among these four proteins are the three histidine kinases present in this proteome AF0103 (HisK-1), AF1332 (HisK-2) and AF1452 (HisK-3). Similar results are obtained if we compare the removal of the proteins ranked eighth and ninth by the two indices. While the removal of those proteins ranked by $EE^0(i)$ has no further changes in the topology of the network the deletion of those ranked by $EE^3(i)$ produce the disconnection of two other proteins, AF0225 and AF0130, respectively.

We consider now the indices $EE^{-t}(i)$ and compare them with the index $EE^0(i)$. The first three entries of the ranking using these indices are the same as the one using $EE^0(i)$. However, the order of the other proteins in the top 10 list change respect to the original index. The most significant change is that protein AF0780/RPA-36 (replication protein A) jumps from the 10th place in the original ranking to the 4th when $t=-4$. For this value of t the protein AF1558/SMC-1 (chromosome segregation protein) has moved from the 13th position in the original ranking to the 10th. These two proteins only have degree two, but they are indeed connected to the hubs of the network, i.e., AF0335 and AF0382. Consequently, these two proteins AF0780/RPA-36 and AF1558/SMC-1 are present in a large number of fragments, which include the two hubs of the network. These fragments are in general large, but the negative rescaling approach does not penalize them very much, which implies that their participation in the centrality is significant enough.

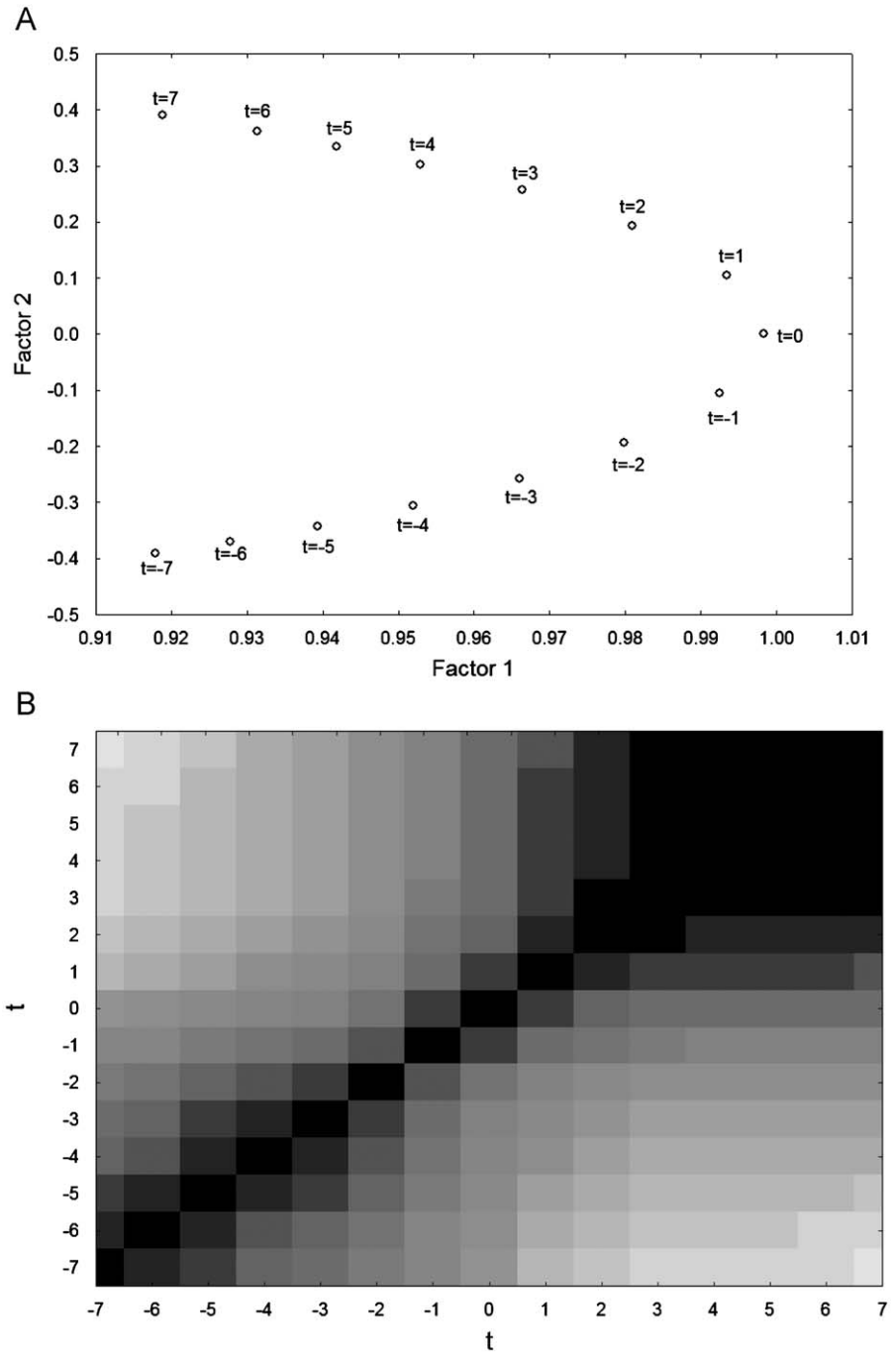

Fig. 3. Statistical analysis of the intercorrelation among the generalized subgraph centrality indices EE^t in the PPI network of the worm *C. elegans* for $-7 \leq t \leq 7$. (A) Plot of the first two factors obtained by using factor analysis. (B) Correlation matrix based on the Kendall τ indices in a gray scale, where white represents $\tau = 0.5$ and black represents $\tau = 1.0$.

In summary, we have seen in this example that the subgraph centrality of a node, e.g., protein can be zoomed in and out in order to extract information which differs in the participation of the nodes in the different structures of the network. In this way, the use of positively rescaled indices zooming in the environment of a node by maximizing the participation of a node in small subgraphs, such as 2-paths, due to the heavy penalization of large spectral moments. These small subgraphs can be important for the communication of different parts of the (protein) networks. On the other hand, when negatively rescaled indices are used, a zooming out of this environment takes places because the participation of a node in large subgraphs is not so heavily penalized, which means that potentially important large substructures of the network are considered for the centrality of a node.

Despite the new indices clearly identify new topological features of the proteins in the PPI networks it is necessary to illustrate whether these indices are useful in explaining or predicting network-independent functional data. In order to conduct this experiment we consider the essentiality of proteins in the PPI network of *S. cerevisiae*. The essentiality of a protein defines the functional significance of a gene at its most basic level. Essential genes are those upon which the cell depends for its viability. Using the GENECENSUS database (http://bioinfo.mbb.

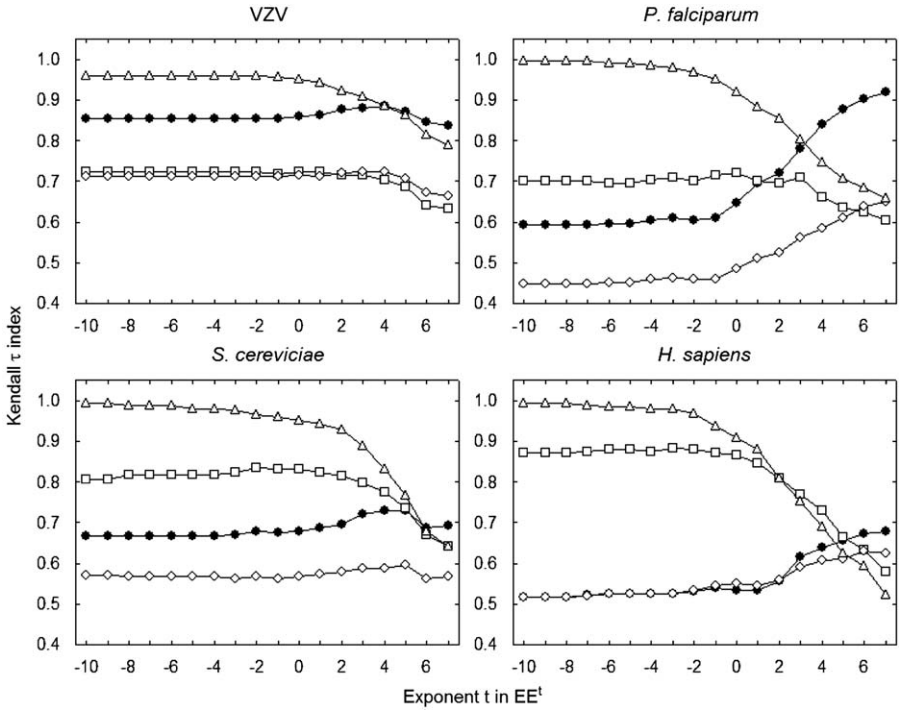

Fig. 4. Plot of the Kendall τ indices for the non-parametric correlation between the generalized subgraph centrality indices EE^t $(-10 \leq t \leq 7)$ and four other centrality measures (DC: filled circles, CC: squares, BC: diamonds, EC: triangles) for four PPI networks.

yale.edu/genome/), we checked for all proteins in the main cluster of the yeast PPI network for essentiality. An illustration of this PPI network and the essential proteins in it is given in Fig. 4.

In a previous work we compared the performance of several centrality measures in identifying essential proteins in this version of the yeast PPI network (Estrada, 2006a). Our strategy consisted in selecting the top top 1%, top 5%, etc., of proteins, and determining how many of these are essential in the yeast PPI network. We showed that the subgraph centrality identifies the largest percentage of essential proteins in comparison with the degree, betweenness, closeness, eigenvector and information centrality. For instance for the top 5% of proteins selected the subgraph centrality identifies 56.4% of essential proteins while a random selection identifies only 25.3% and the degree centrality identifies 41.8%.

Here we use an identical strategy as in our previous work by ranking the proteins in the yeast PPI network according to their values of the walk-based centrality measures introduced here. Then, we select the top 5%, 10%, 15% and 20% and count the number of essential proteins according to each rank. We take the difference between the number of proteins identified by the new indices to the one identified by the subgraph centrality $EE^0(i)$ as an indicator of the performance of the new indices. The indices $EE^{-t}(i)$, which are strongly correlated to $EE^0(i)$, does not show any difference in the number of essential proteins identified in comparison with the subgraph centrality $EE^0(i)$. However, for the positively rescaled indices the number of essential proteins identified is significantly larger than the ones identified by $EE^0(i)$. In Fig. 5 we illustrate the number of essential proteins identified by these indices in excess of that identified by the subgraph centrality. For instance, for the top 5% of proteins selected the index $EE^7(i)$ identifies five essential proteins more than $EE^0(i)$. This index identifies 9, 12 and 17 essential proteins more than $EE^0(i)$ for the top 10%, 15% and 20% of proteins, respectively. This means that $EE^7(i)$ systematically identifies about 4% more essential proteins than the subgraph centrality. Consequently, $EE^7(i)$ outperforms all other centrality indices (degree, betweenness, closeness, eigenvector, information, subgraph) in identifying essential proteins in the yeast PPI network studied here (Fig. 6).

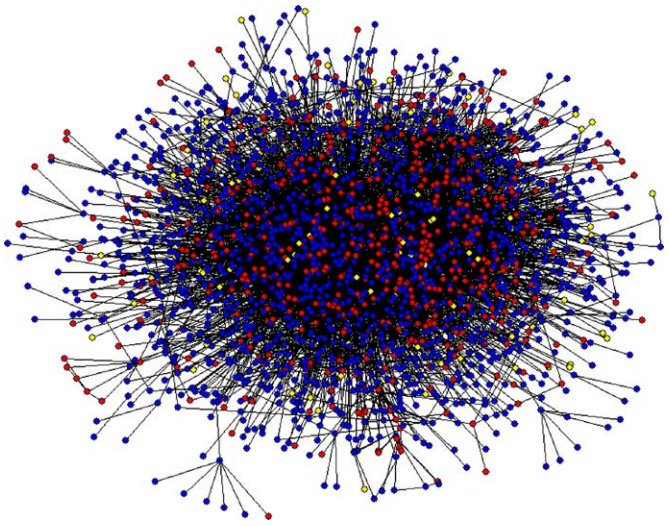

Fig. 5. The principal connected component of the yeast PPI studied here. Red circles represent essential proteins, blue circles represent non-essential ones and yellow circles those whose essentiality is unknown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

8. Conclusions

We have developed a general strategy for zooming in and out the topological environment of a node using a walk-based description of complex networks. Using this approach we generalize the subgraph centrality of nodes in complex networks,

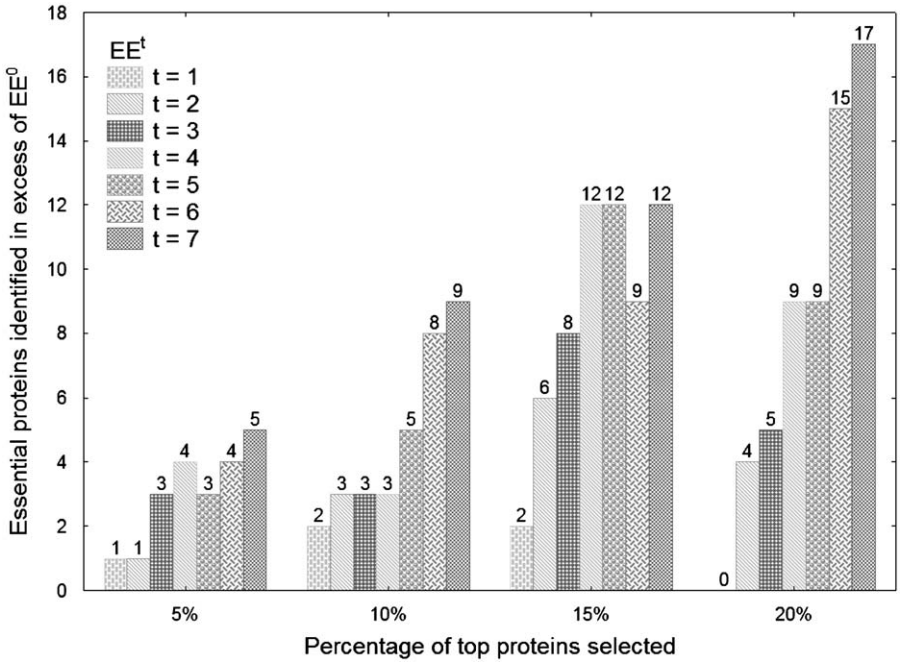

Fig. 6. The number of essential proteins identified by the generalized subgraph centrality indices obtained by positive rescaling EE^t $(1 \leq t \leq 7)$ in excess of those identified by EE^0. The essential proteins are identified among the top 5%, 10%, 15% and 20% of all 2224 proteins in the yeast PPI network.

which is then applied to study PPI networks. The zooming out strategy gives a more global picture of the topological surrounds of a node while the zooming in strategy focuses more on the local topological environment of a node. Subgraph centrality indices based on the last strategy have been able to identify more essential proteins in the yeast PPI network than any of the other centrality measures studied. An important characteristic of these generalized subgraph centrality indices is that we can modulate the zoom around a node to account for more local or global scales of its topological environment. These indices in some way capture the idea of describing some meso-scale environment around a node in which neither very local nor very global environments are accounted for.

Acknowledgments

The author thanks Prof. N.J. Higham (Manchester) for introducing him to the Psi matrix functions. We also thank the comments and suggestions of two anonymous referees, which contributed to a better presentation of the results in this paper. This work is partially supported by the Principal of the University of Strathclyde through the New Professor's Fund.

References

Barabási, A.-L., Oltvai, Z., 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.
Bavelas, A., 1948. A mathematical model for group structure. Hum. Organ. 7, 16–30.
Bavelas, A., 1950. Communication patterns in task oriented groups. J. Acoust. Soc. Am. 22, 271–288.
Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. 2, 113–120.
Bonacich, P., 1987. Power and centrality: a family of measures. Am. J. Sociol. 92, 1170–1182.
Borgatti, S.P., Everett, M.G., 2006. A graph-theoretic perspective on centrality. Soc. Networks 28, 466–484.
Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G., Chen, R., 2003. Topological structure analysis of the yeast protein–protein interaction network of budding yeast. Nucleic Acids Res. 31, 2443–2450.
Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., Emili, A., 2005. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature 433, 531–537.
Carbó-Dorca, R., 2008. Smooth function topological structure descriptors based on graph-spectra. J. Math. Chem. 44, 373–378.
Costa, L.da.F., Rodrigues, F.A., Travieso, G., Boas, P.R.V., 2007. Characterization of complex networks: a survey of measurements. Adv. Phys. 56, 167–242.
de la Peña, J.A., Gutman, I., Rada, J., 2007. Estimating the Estrada index. Lin. Algebra Appl. 427, 70–76.
Deng, H., Radenković, S., Gutman, I., 2009. The Estrada index. In: Cvetković, D., Gutman, I. (Eds.), Applications of Graph Spectra. Mathematical Institute SANU, Beograd.
Estrada, E., 2000. Characterization of 3D molecular structure. Chem. Phys. Lett. 319, 713–718.
Estrada, E., 2002. Characterization of the folding degree of proteins. Bioinformatics 18, 697–704.
Estrada, E., 2004. Characterisation of the amino-acids contributions to the folding degree of proteins. Proteins Struct. Funct. Bioinf. 54, 727–737.
Estrada, E., 2006a. Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics 6, 35–40.
Estrada, E., 2006b. Protein bipartivity and essentiality in the yeast protein–protein interaction network. J. Proteome Res. 5, 2177–2184.
Estrada, E., Hatano, H., 2007. Statistical–mechanical approach to subgraph centrality in complex networks. Chem. Phys. Lett. 439, 247–251.
Estrada, E., Higham, D.H., 2008. Network properties revealed through matrix functions, University of Strathclyde Mathematics Research Report, #17.
Estrada, E., Rodríguez-Velázquez, J.A., 2005. Subgraph centrality in complex networks. Phys. Rev. E 71, 056103.
Estrada, E., Uriarte, E., 2001. Recent advances on the role of topological indices in drug discovery research. Curr. Med. Chem. 8, 1699–1714.
Freeman, L.C., 1979. Centrality networks: I. Conceptual clarifications. Soc. Networks 1, 215–239.
Gursoy, A., Keskin, O., Nussinov, R., 2008. Topological properties of protein interaction networks from a structural perspective. Biochem. Soc. Trans. 36, 1398–1403.
Higham, N.J., 2008. Function of Matrices. Theory and Computation. SIAM, Philadelphia.
Hochbruck, M., Lubich, C., 1999. Exponential integrators for quantum-classical molecular dynamics. BIT 39, 620–645.
Hochbruck, M., Lubich, C., Selhofer, H., 1998. Exponential integrators for large systems of differential equations. SIAM J. Sci. Comput. 19, 1552–1574.
Hoebeke, M., Chiapello, H., Noirot, P., Bessieres, P., 2001. SPiD: a subtilis protein interaction dataset. Bioinformatics 17, 1209–1212.
Jordán, F., Scheuring, I., 2002. Searching for keystones in ecological networks. Oikos 99, 607–612.
Jordán, F., Scheuring, I., 2004. Network ecology: topological constraints on ecosystems dynamics. Phys. Life Rev. 1, 139–172.
Jordán, F., Liu, W.-C., Davis, A.J., 2006. Topological keystone species: measures of positional importance in food webs. Oikos 112, 535–546.

Jordán, F., Benedek, Zs., Podani, J., 2007. Quantifying positional importance in food webs: a comparison of centrality indices. Ecol. Modell. 205, 270–275.

Junker, B.H., Koschützki, D., Schreiber, F., 2006. Exploration of biological network centralities with CentiBiN. BMC Bioinf. 7, 219.

Katz, L., 1953. A new index derived from sociometric data analysis. Psychometrika 18, 39–43.

Kendall, M., 1938. A new measure of rank correlation. Biometrika 30, 81–89.

LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R., Schoenfeld, L.W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., Hughes, R.E., 2005. A protein interaction network of the malaria parasite *Plasmodium falsiparum*. Nature 438, 103–107.

Leavitt, J.J., 1951. Some effects of certain communication patterns on group performance. J. Abnorm. Soc. Psychol. 46, 38–50.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E., Vidal, M., 2004. A map of the interactome network of the metazoan *C. elegans*. Science 303, 540–543.

Lin, C.-Y., Chin, C-H., Wu, H.-H., Chen, S.-H., Ho, C.-W., Ko, M.-T., 2008. Hubba: hubs objects analyzer—a framework of interactome hubs identification for network biology. Nucleic Acids Res. 36, W438–W443.

Motz, M., Kober, I., Girardot, C., Loeser, E., Bauer, U., Albers, M., Moeckel, G., Minch, E., Voss, H., Kilger, C., Koegl, M., 2002. Elucidation of an archaeal replication protein network to generate enhanced PCR enzymes. J. Biol. Chem. 277, 16179–16188.

Noirot, P., Noirot-Gros, M.-F., 2004. Protein interaction networks in bacteria. Curr. Op. Microbiol. 7, 505–512.

Platzer, A., Perco, P., Lukas, A., Mayer, B., 2007. Characterization of protein-interaction networks in tumors. BMC Bioinf. 8, 224.

Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A., Legrain, P., 2001. The protein–protein interaction map of *Helicobacter pylori*. Nature 409, 211–215.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M., 2005. Towards a proteome-scale map of the human protein–protein interaction network. Nature 437, 1173–1178.

Uetz, P., Dong, Y.A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S.V., Roupelieva, M., Rose, D., Fossum, E., Haas, J., 2006. Science 311, 239–242.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Field, S., Bork, P., 2002. Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417, 399–403.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, UK.

Zotenko, E., Mestre, J., O'Leary, D.P., Przytycka, T.M., 2008. Why do hubs in the yeast protein interaction network tend to be essential: re-examining the connection between the network topology and essentiality. PLoS Comput. Biol. 4, 1–16.