A game theoretic neighbourhood-based relevance index

Giulia Cesari, Encarnación Algaba, Stefano Moretti, Juan A. Nepomuceno

Abstract Centrality measures are used in network analysis to identify the relevant elements in a network. Recently, several centrality measures based on coalitional game theory have been successfully applied to different kinds of biological networks, such as *brain networks*, *gene networks*, and *metabolic networks*. We propose an approach, using coalitional games, to the problem of identifying relevant genes in a biological network. Our model generalizes the notion of degree centrality, whose correlation with the relevance of genes for different biological functions is supported by several practical evidences in the literature. The new relevance index we propose is characterized by a set of axioms defined on gene networks and a formula for its computation is provided. Furthermore, an application to the analysis of a large co-expression network is shortly presented.

1 Introduction

Gene regulatory networks and *co-expression networks* are of great interest in the field of molecular biology and epidemiology to better understand the interaction mechanisms between genes, proteins and other molecules within a cell and under certain biological conditions of interest ([5], [7], [8], [34]). A crucial point in the analysis of genes' interaction is the formulation of appropriate measures of the role

Giulia Cesari
Dep. of Mathematics, Politecnico di Milano, Italy e-mail: giulia.cesari@polimi.it

Encarnación Algaba
Dep. of Applied Mathematics and IMUS, Univ. of Seville, Spain. e-mail: ealgaba@us.es

Stefano Moretti
Univ. Paris Dauphine, PSL Research Univ., CNRS, France. e-mail: stefano.moretti@dauphine.fr

Juan A. Nepomuceno
Dep. of Computer Languages and Systems, Univ. of Seville, Spain. e-mail: janepo@us.es

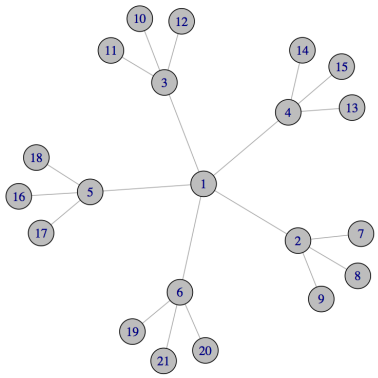
played by each gene to influence the very complex system of genes' relationships in a network.

In our work, we will focus on a particular kind of networks, that are gene co-expression networks, but our approach may be used to analyse other kinds of networks, such as protein-protein interaction networks or cell-cell interaction networks. Co-expression networks [37] may be built from gene expression data collected by means of microarray technology and other high-throughput experimental techniques [26], which allows for the simultaneous quantification of the expression of thousands of genes. The nodes in the co-expression network represent genes (or proteins) and their connection is determined by the coexpression of the genes in the data samples, often measured by the Pearson correlation coefficient between gene expression profiles. This assumption is called the guilt-by-association heuristic: if two genes show similar expression profiles, they are supposed to follow the same regulatory regime. Since the coordinated co-expression of genes encode interacting proteins, studying co-expression patterns can provide insight into the underlying cellular processes and enable the reconstruction of gene regulatory networks.

Centrality analysis represents an important tool for the interpretation of the interaction of genes in a co-expression network ([3], [6], [13],[14], [15]). The relationship between centrality of genes or proteins in a co-expression network and their relevance (measured by biological features such as *lethality* or *essentiality*) has been stressed in several works in the literature. Most central elements of protein networks have been found to be essential to predict lethal mutations [14]. Highly connected hub genes, largely responsible for maintaining network connectivity, have been discovered to be likely essential for yeast survival [6]. In [13] it has been shown how betweenness centrality ([1], [11]) is generally a positive marker for essential genes in *A. thaliana*. Similarly, the relationship between the degree centrality ([25], [32]) and the *essentiality* of genes in transcript co-expression networks has been higlighted in [3]. Moreover, other centrality measures have been investigated in this sense in the recent literature [15].

However, in some cases, genes that lie in the periphery of a network might have an important role in the biological condition it represents. As an example, in [12] it has been shown that differentially expressed genes in major depression (i.e. those genes that present a statistically different behaviour in depressed patients compared to healthy patients) reside in the periphery of resilient gene co-expression networks, thus suggesting that the hub genes are not always the most relevant in the regulatory processes within gene networks. Consider, for instance, the network in Fig 1. All classical centrality measures assign the highest relevance to the hub of the graph, i.e. node 1. Such a node has maximum degree, is the closest node to all other nodes in the graph, lies on the highest number of shortest paths connecting the other nodes and is directly connected with the most nodes of high degree. These features correspond to four of the most known classical centrality [18] measures: *degree centrality* ([25], [32]), *closeness centrality* ([2], [27]), *betweenness centrality* ([1], [11]) and *eigenvector centrality* [4], which give highest centrality to node 1. In particular, the set of nodes $\{2,\ldots,6\}$ has two characteristics that make them relevant genes when the network depicted in Fig 1 represents a gene regulatory network:

Fig. 1 A network with 21 nodes.



(a) through their connections, the nodes in the set are able to influence the expression of all other genes in the network, i.e. they interact directly with all the other genes within the network;
(b) its removal (or inhibition) breaks down the regulatory activity of the network, by leaving all the leaf nodes isolated and therefore not able to maintain their regulatory activity.

With these two features in mind, we introduce an index that aims at measuring the potential of a gene in preserving the regulatory activity within a gene network, by stressing the ability of a gene in influencing the overall expression of genes in the network and to absorb the effects of the inhibition of one or more correlated genes, or in another words its resilience to the removal of connected nodes. In this sense, node 2 (as well as nodes 3,4, 5 and 6) is more relevant than node 1: when node 1 is removed, the network is divided into five components, whose overall regulation is maintained thanks to the presence of nodes 2,3,4,5 and 6 respectively. On the other hand, when one of these last nodes is removed, the network is split in four component, three of whom are no longer able (as being isolated nodes) to maintain their regulatory activity.

The index we propose aims at highlighting the role of genes in the overall "connectivity" of the network, by taking into account the effects that their inhibition have over the induced subnetworks.
A relevant set of genes to this extent would be able to interact directly with the maximum number of other nodes in the network and its removal would split the network in a maximum number of connected components with few genes, or eventually constituted by isolated genes.

To this purpose, we introduce in this paper a *cooperative game*, where the value of a coalition of genes depends on the cardinality of the coalition itself and of its

neighbourhood. The more the genes that are directly interacting in the network with genes in the coalition, and therefore the ability of the coalition to keep the network connected, the higher the power of the coalition. Recently, several centrality measures based on cooperative games have been successfully applied to different kinds of biological networks, such as *brain networks* ([16],[17],[19]), *gene networks* [23], and *metabolic networks* [29], and classical solutions for cooperative games have been employed for the analysis of different biological data ([30],[10],[23]). We introduce our game theoretic relevance index by an axiomatic characterization on gene networks and we provide a formula for its computation, which has a straightforward interpretation. We prove that such an index coincides with the *Shapley value*[31] of the considered cooperative game, which takes into account the marginal contributions of genes to the connectivity of all the coalitions of genes in the network. Moreover, we use our index to assess the relevance of genes in a real dataset related to lung cancer. On such a network, when no *a priori* knowledge is assumed about the genes under analysis, the index is able to highlight the role of genes in the overall connectivity of the network, by assigning the highest relevance to those genes that share the two aforementioned characteristics.

The paper is structured as follows: Section 2 introduces some notations and basic concepts on graph theory and coalitional games. In Section 3 we introduce the methodology, describing our model and an axiomatic characterization of the game-theoretical relevance index in terms of biological properties. An application to gene expression data from microarray technology is presented in Section 4 and Section 5 concludes the paper.

2 Preliminaries and Notations

An (undirected) *graph* or *network* is a pair $\langle N,E\rangle$, where N is a finite set of *vertices* or *nodes* and E is a set of edges e of the form $\{i,j\}$ with $i,j\in N$, $i\neq j$.

We define the set of *neighbours* of a node i in graph $\langle N,E\rangle$ as the set $N_i(E)=\{j\in N:\{i,j\}\in E\}$, and the *degree* of i as the number $d_i(E)=|N_i(E)|$ of neighbours of i in graph $\langle N,E\rangle$. With a slight abuse of notation, we denote by $N_S(E)=\{j\in N:\exists i\in S \text{ s.t. } j\in N_i(E)\}$ the set of neighbours of nodes in $S\in 2^N$, $S\neq\emptyset$, and in the graph $\langle N,E\rangle$. A *path* between nodes i and j in a graph $\langle N,E\rangle$ is a finite sequence of nodes $(i_0,i_1,...,i_k)$, where $i=i_0$ and $j=i_k$, $k\geq 1$, such that $\{i_s,i_{s+1}\}\in E$ for each $s\in\{0,\cdots,k-1\}$ and such that all these edges are distinct. Two nodes $i,j\in N$ are *connected* in $\langle N,E\rangle$ if $i=j$ or if there exists a path between i and j in E. Let $i\in N$ and $S\subseteq N\setminus\{i\}$. A graph $\langle N,E^i_S\rangle$, where the set of edges is $E^i_S=\{\{i,j\}:j\in S\}$ is said a *star on S with center in i*. Notice that the set of neighbours of nodes in $\langle N,E^i_S\rangle$ are such that $N_i(E^i_S)=S$, $N_j(E^i_S)=\{i\}$, for each $j\in S$, and $N_j(E^i_S)=\emptyset$, for each $j\in N\setminus(S\cup\{i\})$.

A *coalitional game* (also known as *cooperative game in characteristic function form* or *Transferable Utility (TU) game*), is a pair (N,v), where N denotes a finite set of *players* and v is the *characteristic function*, assigning to each $S\subseteq N$, a real

number $v(S) \in \mathbb{R}$, with $v(\emptyset) = 0$ by convention. If the set N of players is fixed, we identify a coalitional game (N, v) with the corresponding characteristic function v. A group of players $S \subseteq N$ is called a *coalition* and $v(S)$ is called the *worth* of coalition S. We will denote by $\mathscr{G}$ the class of all coalitional games. Let $\mathscr{C} \subseteq \mathscr{G}$ be a subclass of coalitional games. Given a set of players N, we denote by $\mathscr{C}^N \subseteq \mathscr{C}$ the class of coalitional games in $\mathscr{C}$ with N as set of players. A *one-point solution* (or simply a *solution*) for a class $\mathscr{C}^N$ of coalitional games is a function $\psi : \mathscr{C}^N \to \mathbb{R}^N$ that assigns a payoff vector $\psi(v) \in \mathbb{R}^N$ to every coalitional game in the class. It prescribes how to convert the information on the worth of every coalition of players in a single attribution to each of the players. A well-known solution is the Shapley value [31], which has been applied to a wide range of fields, including biology [24]. The Shapley value $\phi_i(v)$ of a player $i \in N$ in a game (N, v) is defined as the *average marginal contribution* of i over all $|N|!$ possible orders of players (we denote by $|N|$ the cardinality of the set N), and can be computed according to the following formula:

$$\phi_i(v) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(v(S \cup i) - v(S)). \tag{1}$$

We recall some nice properties of the Shapley value of a coalitional game (N, v) [31]: *efficiency* (EFF), i.e. $\sum_{i \in N} \phi_i(v) = v(N)$; *symmetry* (SYM), i.e. if $i, j \in N$ are such that $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$, then $\phi_i(v) = \phi_j(v)$; *dummy player property* (DPP), i.e. if $i \in N$ is such that $v(S \cup \{i\}) - v(S) = v(\{i\})$ for all $S \subseteq N$, then $\phi_i(v) = v(\{i\})$; *additivity* (ADD), i.e. $\phi(v) + \phi(w) = \phi(v + w)$ for each $v, w \in \mathscr{C}^N$. It is well known that the Shapley value is the only solution that satisfies these four properties on the class $\mathscr{G}^N$.

3 Methodology

Let $\langle N, E \rangle$ be a *gene network*, that is a network where the set of nodes N represents a set of genes and the set of edges E describes the interaction among genes, i.e. there exists an edge between two genes if they are directly interacting in the biological condition under analysis. Moreover, let $k \in \mathbb{R}^N$ be a parameter vector that specifies the *a priori* importance of each gene. We define a coalitional game (N, v_E^k), where N is the set of genes under study and the characteristic function v_E^k assigns a worth to each coalition of genes $S \subseteq N$ representing the overall magnitude of the interaction between the genes in S, which takes into account the weight (*a priori* importance) of each gene directly connected to S in the biological network.
More precisely, the map $v_E^k : 2^N \to \mathbb{R}$ assigns to each coalition $S \in 2^N \setminus \{\emptyset\}$ the value

$$v_E^k(S) = \sum_{j \in S \cup N_S(E)} k_j \tag{2}$$

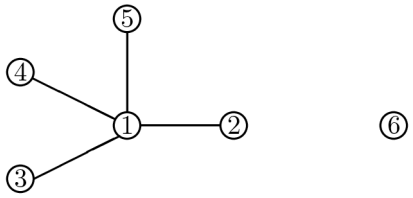


Fig. 2 A star with six nodes. The network in figure is the star $\langle\{1,2,3,4,5,6\},E^1_{\{2,3,4,5\}}\rangle$.

that is the sum of the weights associated to the genes in S and to the ones that are directly connected in $\langle N,E\rangle$ to some genes in S (notice that v^k_E specifies a real number for each of the $2^{|N|}$ subsets of N and, by convention, the value of the empty coalition is null, i.e., $v^k_E(\emptyset)=0$). The class of games (N,v) defined according to relation (2), on some gene network $G\equiv\langle V,E\rangle$ and with parameter $k\in\mathbb{R}^N$, is denoted by $\mathscr{EK}^N$.

In the literature related to the application of game theoretic centrality to co-expression networks, another way to keep into account the *a priori* importance of genes has been proposed in [23] by means of the so-called *association game*, where a set of key-genes $K\subset N$ (e.g. a set of genes known *a priori* to be involved in biological pathways related to chromosome damage) is considered and the value assigned to a coalition S is the number of key-genes interacting only with S (formally, in [23] the value assigned to a coalition $S\subseteq N$ is the cardinality of the set $\{i\in K : N_i(E)\subseteq S\}$). However, the definition proposed in relation (2) is more flexible to explore all possibilities of reciprocal influence among genes. It generalises the game introduced in [35] for determining the "top-k nodes" in a co-autorship network, by the introduction of a parameter that specifies the *a priori* importance of each node. The parameter vector k allows for an *a priori* ranking of the genes according to their importance, while in the previous model introduced in [23] only a two-level distinction was made between key-genes and non key-genes. Moreover, by measuring to what extent a coalition of genes is connected to the rest of the network, relation (2) generalizes the notion of degree centrality for groups of genes, which is justified by some practical evidences showing a strong correlation between the degree centrality and genes that are essential for different biological functions (see, for instance, [3, 6, 14, 15]).

We now introduce some properties for a *relevance index* for genes, that is a map $\rho:\mathscr{EK}^N\to\mathbb{R}^N$. We start with a reinterpretation of the classical properties of SYM, DPP and EFF on the class $\mathscr{EK}^N$ (see Section 2 for a formal definition on the class of all TU-games).

Consider a gene network $\langle N,E\rangle$ and a vector of weights $k\in\mathbb{R}^N$. The property of SYM implies that if two genes $i,j\in N$ have the same weight ($k_i=k_j$) and in addition, they are connected to the same set of neighbours ($N_i(E)=N_j(E)$), then they should have the same relevance. For instance, nodes $2,3,4$ and 5 in the star depicted in Fig 2 are symmetric.

The property DPP also has an intuitive interpretation on the graph: every disconnected node $i \in N$ (like node 6 in Fig 2) should have relevance k_i. Finally, the EFF property implies that the sum of the relevance of all genes should be equal to $\sum_{i \in N} k_i$, the total sum of weights.

We introduce now a new axiom, saying that the transformation of a node i with zero weight to a node with weight k_i should affect only the genes directly connected to i, and its impact on the relevance of its neighbours should be equal to the one had in an equivalent star of center i.

Axiom 1 (Star Additivity, SADD) *Let $\langle N, E \rangle$ be a gene network with parameter vector $k_{-i} \in \mathbb{R}^N$ such that gene i has weight 0 and let $v_E^{k_{-i}}$ be the corresponding game defined according to relation (2). Then consider the game v_E^k defined according to relation (2) on $\langle N, E \rangle$ and with parameter vector k that assigns a positive weight k_i to gene i and the same weight as k_{-i} to all the other genes. An index $\rho : \mathscr{EK}^N \to \mathbb{R}^N$ satisfies the SADD property iff*

$$\rho(v_E^k) = \rho(v_E^{k_{-i}}) + \rho(v^{s^i}_{E^i_{N_i(E)}}),$$

where $v^{s_i}_{E^i_{N_i(E)}}$ is the game defined according to relation (2) on the star $\langle N, E^i_{N_i(E)} \rangle$ on $N_i(E)$ with center i and s^i is the parameter vector that assigns k_i to i and 0 to $j \neq i$.

For instance, consider again the network of Fig 2, and suppose that $\rho' \in \mathbb{R}^6$ is the relevance index corresponding to a parameter vector k_{-1}. Moreover let $\rho'' \in \mathbb{R}^6$ be the relevance index on the same network with parameter s^1 such that only node 1 has a positive weight k_1. Then, the SADD property says that in the situation where the parameter vector is given by $k = k_{-1} + s^1$ and $\rho''' \in \mathbb{R}^6$ is the corresponding relevance index, then it must hold $\rho''' = \rho' + \rho''$.
Roughly speaking, axiom SADD states that increasing the weight of a node i from 0 to a positive value should only affect the total relevance of gene i and its neighbours at the same extent for whatever graph. As a consequence, a positive change in the weight of a gene produces the same effect on its relevance and on the one of their neighbours independently from the topology of the network, and the effect of the changes is comparable along different networks. Then, redistributing the *a priori* importance of a node among their neighbours, the SADD property catches the idea of measuring the capacity of nodes to absorb the effect of inhibition of correlated genes, as previously discussed in Section 1.

Proposition 1. *The Shapley value is the unique relevance index ρ that satisfies SYM, DPP, EFF and SADD on the class $\mathscr{EK}^N$. Moreover, for each gene network $\langle N, E \rangle$ with $k \in \mathbb{R}^N$ as a vector of weights, it can be computed according to the following formula:*

$$\rho_i(v_E^k) = \sum_{j \in (N_i(E) \cup \{i\})} \frac{k_j}{d_j(E)+1}, \tag{3}$$

for each $i \in N$.

Proof. Let $\langle N, E^i_{N_i(E)} \rangle$ be a star on $N_i(E)$ with center in i, and such that only i has a positive weight equal to k_i and let $v^{s^i}_{E^i_{N_i(E)}}$ be the corresponding game defined according to relation (2). It is easy to check that the unique index that satisfies the properties of SYM, DPP and EFF is the one such that

$$\rho_j(v^{s^i}_{E^i_{N_i(E)}}) = \phi_j(v^{s^i}_{E^i_{N_i(E)}}) = \begin{cases} \frac{k_i}{d_i(E)+1} & \text{if } j \in N_i(E) \cup \{i\}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

By the repeated application of axiom SADD, and since $\sum_{i \in N} v^{s^i}_{E^i_{N_i(E)}} = v^k_E$, we have that

$$\rho(v^k_E) = \sum_{i \in N} \rho(v^{s^i}_{E^i_{N_i(E)}}). \tag{5}$$

Then, the proof follows by relation (4) and the additivity of the Shapley value.

The interpretation of the formula in (3) is straightforward: a gene is assigned a high relevance if it is connected to many genes which are in turn connected with few other genes, that is the more neighbours with low degree, the highest the relevance. Notice also that, due to the exponential number of terms involved, the Shapley value is general computationally challenging. Instead, for the class of games considered in this section, formula (3) allows for the computation of the Shapley value in polynomial time.

Example 1. Consider the gene network in Fig 1. Suppose all the genes have the same *a priori* importance, and let for simplicity $k_i = 1 \ \forall i \in N$. Then, by Proposition 1 $\rho(v^k_E) = (\frac{35}{30}, \frac{56}{30}, \frac{56}{30}, \frac{56}{30}, \frac{56}{30}, \frac{56}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30}, \frac{21}{30})$.

Therefore, our index gives the highest relevance to nodes $2,3,4,5$ and 6, followed by node 1 and the least relevance to the leaf nodes $\{7,\ldots,21\}$. On the other hand, all the other classical centrality measures defined in Section 2 provide the following ranking: node 1 has the maximum centrality, followed by nodes $\{2,3,4,5,6\}$ and finally the leaf nodes.

4 Experimental results

A gene expression dataset related with a very common kind of lung cancer called adenocarcinoma has been studied. Adenocarcinoma cancers are usually found in lung outer areas as the lining of the airways. The data were generated in a study where 107 samples of several tumor stages in a population of smoker and not smoker people were analyzed [20] (the dataset with accession number GDS3257 has been downloaded from the Gene Expression Omnibus (GEO) [9]). These raw data have been preprocessed with Babelomics tool [22] using some standard filtering steps. Concretely, those genes with a percentage of missing values greater than 80% have

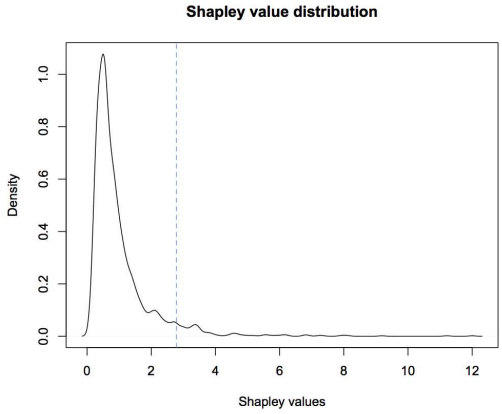


Fig. 3 First analysis: Shapley value distribution. The density distribution of the index ρ is shown, for $k_i = 1$ for every gene i. The dotted vertical line represents the cutoff: the 5% of genes with highest index is selected.

been removed. In the rest of the cases, missing values have been replaced with the average of the expression profile of the row. Those gene profiles with a standard deviation under 0.5 have been removed in order to only consider genes differentially expressed. The resulting gene expression matrix is composed by 2517 gene expression profiles (rows) and 107 samples (columns).

A gene co-expression network has been generated, by establishing a link between two genes if and only if the Pearson's correlation between their gene expression profiles is higher than a fixed threshold. The choice of the threshold is based on the following considerations: a suitable network should consist of connected components with the highest possible cardinality and should also be as sparse as possible in order to better reveal the relationships between the nodes (genes). Therefore, the network must be experimentally built according to an equilibrium between connectivity and sparsification [33]. The BioLayout tool [36] has been used to conduct an experimental study, which has led to the choice of 0.8 as the value for the correlation threshold. The network so obtained is composed by 2154 nodes (genes) and 24821 edges.

A first analysis has been carried out on the aforementioned network, with no *a priori* knowledge of the importance of the different genes, thus considering each gene equally important, i.e. setting $k_i = 1$ for each gene $i \in N$. Following this approach, the relevance index ρ is computed. The density distribution of ρ is shown in Fig 3. In particular, we select the 5% of genes with highest relevance for further analysis ($n = 108$). The lists of the 5% of genes with highest value according to the different centrality measures are compared as follows:

(i) the 108 genes selected by our index are directly interacting in the network with 1412 genes, comparably with the ones selected by the betweenness centrality, whose neighborhood consists in 1423 genes. The other measures are much less effective in this sense: the genes selected by the degree centrality interact with 1062 genes, the ones by closeness centrality with 668 and the ones by eigenvector centrality with 383 genes.

(ii)when the 108 genes selected by ρ are removed, the network is split in 165 connected components, 125 of which are isolated nodes. Three of them contain a high number of genes (550, 826 and 338), one of them 42 nodes, and the rest very few nodes (2 to 10 nodes each). A similar behaviour is observed after the removal of the 108 nodes selected by the betweenness centrality: the network is split in 170 components, 122 of which are isolated nodes. On the other hand, the effects of the removal of the genes selected by the other measures are definitively less severe.

Among the classical centrality measures, our relevance index shows a maximum overlap with betweenness centrality, with 66 genes in common (out of the 108 selected) and a high positive correlation between the list of genes. The number of common genes among the different lists and their correlation, measured by the Pearson correlation coefficient, are shown in Table 1. The set of 100 genes with the highest

Table 1 Number of common genes among the relevance vectors of 108 genes provided by the different relevance measures. The number in parenthesis represents the correlation of the vector of indices among the lists of common genes.

	$\rho(1)$	degree	closeness	betweenness	eigenvector
$\rho(1)$	108 (1)	49 (-0.221)	40 (0.430)	66 (0.846)	28 (0.509)
degree	49 (-0.221)	108 (1)	19 (0.482)	28 (-0.068)	86 (0.977)
closeness	40 (0.430)	19 (0.482)	108 (1)	48 (0.578)	0 (NA)
betweenness	66 (0.846)	28 (-0.068)	48 (0.697)	108 (1)	7 (0.121)
eigenvector	28 (0.509)	86 (0.977)	0 (NA)	7 (0.121)	108 (1)

Shapley value has also been investigated from a biological point of view. A Literature Mining approach has been used with a *Cytoscape* plugging called *Agilent Literature Search* [28]. The Cytoscape plugging searches a set of genes in published papers available in public repositories such as *PubMed*. The search has been performed by taking as input the list of genes selected by our relevance index and a set of key-words, namely "Homo sapiens" and "Adenocarcinoma". The tool provides as a result that the subset of selected genes that are cited in the related literature is composed by 70, which is a rather encouraging result. More details about the experimental analysis are omitted for space reasons.

5 Conclusions

In this paper, we proposed a relevance index for nodes of gene co-expression networks, with the objective of measuring the potential of genes in acting as intermediaries between hub nodes and leaf nodes and preserving the regulatory activity within gene networks. For this purpose, we used a game-theoretic approach, by defining a cooperative game where the strength of a coalition of genes depends on the *a priori*

importance of the genes in its neighbourhood. The Shapley value of such a game is proposed as a new relevance index for genes. Our methodology is supported by a property-driven approach, where the set of properties satisfied by our index have a biological interpretation.

The versatility of our model allows for the combination of a game-theoretical approach with other techniques from network analysis in order to assess the *a priori* importance of genes (i.e., the parameter vector k) in the network under analysis, which may be used as a parameter of the model to compute the *a posteriori* relevance of nodes. For instance, in an alternative analysis (omitted for space reasons) of the same lung cancer network considered in Section 4, the *a priori* importance of genes has been assessed by a parameter vector that depends on the cluster structure of the network, and according to the principle that overlapping genes among clusters are to some extent important in a gene network [21]. An interesting direction for future research is to further explore these techniques, in order to refine the relevance analysis, and to compare the results provided by our model over different co-expression networks.

References

1. Bavelas, A.: A mathematical model for group structures. Human organization **7**(3), 16–30 (1948)
2. Beauchamp, M.A.: An improved index of centrality. Systems Research and Behavioral Science **10**(2), 161–163 (1965)
3. Bergmann, S., Ihmels, J., Barkai, N.: Similarities and differences in genome-wide expression data of six organisms. PLoS biology **2**(1), e9 (2003)
4. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology **2**(1), 113–120 (1972)
5. Butte, A.J., Kohane, I.S.: Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Pacific Symposium on Biocomputing, vol. 5, p. 26 (2000)
6. Carlson, M.R., Zhang, B., Fang, Z., Mischel, P.S., Horvath, S., Nelson, S.F.: Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC genomics **7**(1), 40 (2006)
7. Carter, S.L., Brechbühler, C.M., Griffin, M., Bond, A.T.: Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics **20**(14), 2242–2250 (2004)
8. Davidson, E.H., McClay, D.R., Hood, L.: Regulatory gene networks and the properties of the developmental process. Proceedings of the National Academy of Sciences **100**(4), 1475–1480 (2003)
9. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research **30**(1), 207–210 (2002)
10. Fagnocchi, L., Bottini, S., Golfieri, G., Fantappiè, L., Ferlicca, F., Antunes, A., Guadagnuolo, S., Del Tordello, E., Siena, E., Serruto, D., et al.: Global transcriptome analysis reveals small rnas affecting neisseria meningitidis bacteremia. PLoS One **10**(5), e0126,325 (2015)
11. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)
12. Gaiteri, C., Sibille, E.: Differentially expressed genes in major depression reside on the periphery of resilient gene coexpression networks. Frontiers in neuroscience **5** (2011)

13. Giorgi, F.M., Del Fabbro, C., Licausi, F.: Comparative study of rna-seq-and microarray-derived coexpression networks in arabidopsis thaliana. Bioinformatics **29**(6), 717–724 (2013)
14. Jeong, H., Mason, S., Barabási, A.L., Oltvai, Z.: Lethality and centrality in protein networks. Nature **411**(6833), 41 (2001)
15. Junker, B.H., Koschützki, D., Schreiber, F.: Exploration of biological network centralities with centibin. BMC bioinformatics **7**(1), 219 (2006)
16. Kaufman, A., Keinan, A., Meilijson, I., Kupiec, M., Ruppin, E.: Quantitative analysis of genetic and neuronal multi-perturbation experiments. PLoS computational biology **1**(6), e64 (2005)
17. Keinan, A., Sandbank, B., Hilgetag, C.C., Meilijson, I., Ruppin, E.: Fair attribution of functional contribution in artificial and biological networks. Neural Computation **16**(9), 1887–1915 (2004)
18. Koschützki, D., Lehmann, K.A., Peeters, L., Richter, S., Tenfelde-Podehl, D., Zlotowski, O.: Centrality indices. In: Network analysis, pp. 16–61. Springer (2005)
19. Kötter, R., Reid, A.T., Krumnack, A., Wanke, E., Sporns, O.: Shapley ratings in brain networks. Frontiers in neuroinformatics **1** (2007)
20. Landi, M.T., Dracheva, T., Rotunno, M., Figueroa, J.D., Liu, H., Dasgupta, A., Mann, F.E., Fukuoka, J., Hames, M., Bergen, A.W., et al.: Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PloS one **3**(2), e1651 (2008)
21. Li, J., Halgamuge, S.K., Tang, S.L.: Genome classification by gene distribution: an overlapping subspace clustering approach. BMC evolutionary biology **8**(1), 116 (2008)
22. Medina, I., Carbonell, J.., Pulido, L., Madeira, S.C., Goetz, S., Conesa, A., Tï¿ 1/2rraga, J.n., Pascual-Montano, A., Nogales-Cadenas, R., Santoyo, J., et al.: Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic acids research **38**(suppl_2), W210–W213 (2010)
23. Moretti, S., Fragnelli, V., Patrone, F., Bonassi, S.: Using coalitional games on biological networks to measure centrality and power of genes. Bioinformatics **26**(21), 2721–2730 (2010)
24. Moretti, S., Patrone, F.: Transversality of the shapley value. Top **16**(1), 1–41 (2008)
25. Nieminen, J.: On the centrality in a graph. Scandinavian journal of psychology **15**(1), 332–336 (1974)
26. Parmigiani, G., Garrett, E.S., Irizarry, R.A., Zeger, S.L.: The analysis of gene expression data: an overview of methods and software. In: The analysis of gene expression data, pp. 1–45. Springer (2003)
27. Sabidussi, G.: The centrality index of a graph. Psychometrika **31**(4), 581–603 (1966)
28. Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Lotia, S., Pico, A.R., Bader, G.D., Ideker, T.: A travel guide to cytoscape plugins. Nature methods **9**(11), 1069–1076 (2012)
29. Sajitz-Hermstein, M., Nikoloski, Z.: Restricted cooperative games on metabolic networks reveal functionally important reactions. Journal of theoretical biology **314**, 192–203 (2012)
30. Sajitz-Hermstein, M., Nikoloski, Z.: Structural control of metabolic flux. PLoS computational biology **9**(12), e1003,368 (2013)
31. Shapley, L.: A value for n-person games. In: K. H., T. A.W. (eds.) Contributions to the Theory of Games II, pp. 307–317. Princeton University Press (1953)
32. Shaw, M. E.: Group structure and the behavior of individuals in small groups. The Journal of psychology **38**(1), 139–149 (1954)
33. Silva, T.C., Zhao, L.: Machine learning in complex networks, vol. 2016. Springer (2016)
34. Stuart, J. M., Segal, E., Koller, D., Kim, S. K. : A gene-coexpression network for global discovery of conserved genetic modules. Science **302**(5643), 249–255 (2003)
35. Suri, N. R., Narahari, Y.: Determining the top-k nodes in social networks using the shapley value. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, vol. 3, pp. 1–13. Springer Berlin Heidelberg (2010)
36. Theocharidis, A., Van Dongen, S., Enright, A.J., Freeman, T.C.: Network visualization and analysis of gene expression data using biolayout express3d. Nature protocols **4**(10), 1535–1550 (2009)
37. Zhang, B., Horvath, S. : A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology **4**(1), 1128 (2005)