BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 21 2010, pages 2721–2730
doi:10.1093/bioinformatics/btq508

Gene expression

Advance Access publication September 3, 2010

# Using coalitional games on biological networks to measure centrality and power of genes

Stefano Moretti[1,2,*], Vito Fragnelli[3], Fioravante Patrone[4] and Stefano Bonassi[5]

[1]CNRS, FRE3234, [2]Université Paris-Dauphine, Lamsade, F-75016 Paris, France, [3]Department of Advanced Sciences and Technologies, University of Eastern Piedmont, Alessandria, [4]DIPTEM, University of Genoa, Genoa and [5]Unit of Clinical and Molecular Epidemiology, IRCCS San Raffaele Pisana, Rome, Italy

Associate Editor: Trey Ideker

## ABSTRACT

**Motivation:** The interpretation of gene interaction in biological networks generates the need for a meaningful ranking of network elements. Classical centrality analysis ranks network elements according to their importance but may fail to reflect the power of each gene in interaction with the others.

**Results:** We introduce a new approach using coalitional games to evaluate the centrality of genes in networks keeping into account genes' interactions. The Shapley value for coalitional games is used to express the power of each gene in interaction with the others and to stress the centrality of certain hub genes in the regulation of biological pathways of interest. The main improvement of this contribution, with respect to previous applications of game theory to gene expression analysis, consists in a finer resolution of the gene interaction investigated in the model, which is based on pairwise relationships of genes in the network. In addition, the new approach allows for the integration of a priori knowledge about genes playing a key function on a certain biological process. An approximation method for practical computation on large biological networks, together with a comparison with other centrality measures, is also presented.

**Contact:** stefano.moretti@dauphine.fr

Received on September 17, 2009; revised on August 2, 2010; accepted on August 31, 2010

## 1 INTRODUCTION

Gene expression data may be collected by means of microarray technology (Golub *et al.*, 1999; Parmigiani *et al.*, 2003). Within a single experiment of this sophisticated technology, the level of expression of thousands of genes is estimated in a sample of cells under given conditions (genetic diseases, environmental exposition, pharmacologic treatment, levels of activation of a given pathway of genes, etc.). Several approaches have been proposed to identify 'central' genes of different biological pathways within the huge amount of information provided by this technology (Amaratunga and Cabrera, 2004; Storey and Tibshirani, 2003; Tusher *et al.*, 2001).

Gene co-expression networks (Zhang and Horvath, 2005) and other biological networks (e.g. representing protein–protein interactions) are increasingly used to explore the system-level functionality of genes and proteins (Carlson *et al.*, 2006; Jeong *et al.*, 2001). Co-expression networks, for instance, are connection situations based upon the extent of correlation between pairs of genes across a gene expression dataset. Nodes are genes and connections are defined by co-expression of two genes. Often, the Pearson's correlation coefficient is the initial measure of gene co-expression. This measure is then transformed into an adjacency matrix, according to different alternative statistical procedures (Carlson *et al.*, 2006; Zhang and Horvath, 2005). Depending on the aims of the study, weighted or unweighted networks, generated by the dichotomization of the corresponding correlation matrix, may be considered. Analytical methods for network elements ranking are an important tool for the interpretation of gene interaction in co-expression networks. Centrality analysis ranks single elements according to their importance within the network structure, and different measures of centrality focus on various aspects of the structure of a network (Junker *et al.*, 2006; Mason and Verwoerd, 2007), e.g. most central elements of protein networks were essential to predict lethal mutations (Jeong *et al.*, 2001). Highly connected hub genes, largely responsible for maintaining network connectivity, were likely essential for yeast survival (Carlson *et al.*, 2006), although standard centrality measures may fail to reflect the power of each gene to interact with the others.

Cooperative game theory may also be used to analyze gene expression data see, (for instance, Albino *et al.*, 2008; Esteban and Wall, 2009; Lucchetti *et al.*, 2009; Fragnelli and Moretti, 2008; Moretti, 2009, 2010; Moretti *et al.*, 2007, 2008). In (Moretti *et al.*, 2007), the class of microarray games has been introduced to quantitatively evaluate the relevance of each gene in generating or regulating a condition of interest (e.g. a disease), taking into account the observed relationships in all subgroups of genes. In the framework of microarray games, the relevance of genes is expressed in terms of the Shapley value (Moretti and Patrone, 2008; Shapley, 1953). The Shapley value attributed to a certain gene in a given microarray game corresponds to the relevance of that gene for the mechanisms governing the genomic effects of the condition under study. This game-theoretic approach has been successfully applied to real datasets (Albino *et al.*, 2008; Moretti *et al.*, 2008) and provides a characterization of a relevance index for genes which is mainly based on the role they play inside gene-regulatory pathways (Moretti *et al.*, 2007). A comparison between the results provided by the analysis of the Shapley value of microarray games and the results provided by classic statistical testing is discussed in connection with the pathogenesis of neuroblastic tumors in (Albino *et al.*, 2008),

[*]To whom correspondence should be addressed.

and in (Moretti *et al*., 2008), where gene expression in children differentially exposed to air pollution is studied.

Standard centrality measures (Junker *et al*., 2006; Mason and Verwoerd, 2007) do not take into account the strength of interrelations inside subgroups of genes, in contrast with a central issue of coalitional games in cooperative game theory, which is precisely to analyze the overall power of players according to their role in all feasible 'coalitions'. In the context of social networks, (Gòmez *et al*., 2003) proposed a new family of centrality measures based on coalitional games defined on networks. Our idea was to use a similar approach in the context of co-expression networks. We define an association game as a coalitional game (also known as a cooperative game in characteristic function form) $(N,v)$, where $N$ is the set of genes studied in the expression dataset and $v$ is the characteristic function, which assigns a 'worth' to each subset (coalition) of genes in $N$. The worth of a coalition represents the overall magnitude of the correlation between the genes of the coalition and a set of key genes selected a priori (e.g. a set of genes known to be involved in biological pathways related to chromosome damage).

In order to study the cascade of activation/deactivation among genes, gene interaction is restricted to the connections within an associated interaction network or co-expression network $\Gamma$, and therefore another coalitional game $(N,w^v_\Gamma)$ is studied, which is defined as the restriction of the association game $(N,v)$ to the co-expression network computed on the dataset (Myerson, 1977). The difference of the Shapley values computed on the two coalitional games $(N,v)$ and $(N,w^v_\Gamma)$ is considered as a gene centrality measure.

The article is organized as follows. Next section, after the introduction of some preliminary notations, is devoted to the presentation of the game-theoretic centrality measure. Section 3 presents a preliminary application of the method to a real dataset. Section 4 introduces an approximation method for centrality computation and the comparison of the results with other centrality measures on a large network. Section 5 concludes the article.

## 2 APPROACH

### 2.1 Preliminaries

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

A *path* between nodes $i$ and $j$ in a graph $\langle V,E\rangle$ is a finite sequence of nodes $(i_0,i_1,\ldots,\ i_k)$, where $i=i_0$ and $j=i_k$, $k\geq 1$, such that $\{i_s,i_{s+1}\}\in E$ for each $s\in\{0,\ldots,k-1\}$ and such that all these edges are distinct. Two nodes $i,j\in V$ are connected in $\langle V,E\rangle$ if $i=j$ or if there exists a path between $i$ and $j$ in $E$. The *length* of a path between $i$ and $j$ in a graph $\langle V,E\rangle$ is the number of edges in the path and a *shortest path* between $i$ and $j$ in a graph $\langle V,E\rangle$ is a path between $i$ and $j$ with minimum length. We denote by $\mathcal{SP}^{i,j}_E$ the set of all the shortest paths between two nodes $i,j\in V$, $i\neq j$, in a graph $\langle V,E\rangle$.

A *cycle* in $\langle V,E\rangle$ is a path from $i$ to $i$ for some $i\in V$. A path $(i_0,i_1,\ldots,i_k)$ is *without cycles* if there do not exist $a,b\in\{0,1,\ldots,k\}$, $a\neq b$, such that $i_a=i_b$. A *forest* is a graph where each path is without cycles.

A *connected component* of $V$ in $\langle V,E\rangle$ is a maximal subset of $V$ with the property that any two nodes in this subset are connected in $\langle V,E\rangle$. The set of all the connected components in $\langle V,E\rangle$ is denoted by $\mathcal{C}_E$.

Now, we introduce some basic game-theoretical notations. A *coalitional game* or *characteristic-form game* is a pair $(N,v)$, where $N$ denotes a finite set of *players* and $v$ is the *characteristic function*, assigning to each $S\subseteq N$, $v(S)\in\boldsymbol{R}$, 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 $T\subseteq N$ is called a *coalition* and $v(T)$ is called the *worth* of this coalition. We will denote by $\mathcal{G}$ the class of all coalitional games.

Let $\mathcal{C}\subseteq\mathcal{G}$ be a subclass of coalitional games. Given a set of players $N$, we denote by $\mathcal{C}^N\subseteq\mathcal{C}$ the class of coalitional games in $\mathcal{C}$ with $N$ as set of players.

The *unanimity game* $(N,u_R)$ on $\emptyset\neq R\subseteq N$ is the game described by $u_R(T)=1$ if $R\subseteq T$ and $u_R(T)=0$, otherwise. Every coalitional game $(N,v)$ can be written as a linear combination of unanimity games in a unique way, *i.e.* $v=\sum_{S\subseteq N,S\neq\emptyset}\lambda_S(v)u_S$ (see, for instance, Owen, 1995). The coefficients $\lambda_S(v)$, for each $S\in 2^N\setminus\{\emptyset\}$, are called *unanimity coefficients* of the game $(N,v)$.

A *payoff vector* or *allocation* $(x_1,\ldots,x_n)$ of a coalitional game $(N,v)$ is a vector $\in\boldsymbol{R}^N$ describing the payoffs of the players, such that each player $i\in N$ receives $x_i$.

A *one-point solution* (or simply a *solution*) for a class $\mathcal{C}^N$ of coalitional games is a function $\psi$ that assigns a payoff vector $\psi(v)$ to every coalitional game in the class, that is $\psi:\mathcal{C}^N\to\boldsymbol{R}^N$.

The most widely used solution in the theory of coalitional games is the *Shapley value*, introduced in (Shapley, 1953). This solution can be described in several ways. First, we need to introduce the notions of order on $N$ and of marginal vector.

We define the set $\Sigma_N$ of possible orders on the set $N$ as the set of all bijections $\sigma:N\to N$, where $\sigma(i)=j$ means that with respect to $\sigma$, player $j$ is in the $i$-th position. Let $(N,v)$ be a coalitional game with $N$ as the set of players. For $\sigma\in\Sigma_N$, the *marginal vector* $m^\sigma(v)$ is defined by

$$m^\sigma_i(v)=v([i,\sigma])-v((i,\sigma))\text{ for all }i\in N,$$

where $[i,\sigma]=\{j\in N:\sigma^{-1}(j)\leq\sigma^{-1}(i)\}$ is the set of predecessors of $i$ with respect to $\sigma$ including $i$, and $(i,\sigma)=\{j\in N:\sigma^{-1}(j)<\sigma^{-1}(i)\}$ is the set of predecessors of $i$ with respect to $\sigma$ excluding $i$.

The Shapley value $\phi(v)$ of a game $(N,v)$ is then defined as the average of marginal vectors over all $|N|!$ possible orders in $\Sigma_N$ ($|N|$ is the cardinality of the set $N$). In formula

$$\phi_i(v)=\sum_{\sigma\in\Sigma_N}\frac{m^\sigma_i(v)}{|N|!}\text{ for all }i\in N. \tag{1}$$

An alternative representation of the Shapley value can be given in terms of the unanimity coefficients $(\lambda_S(v))_{S\in 2^N\setminus\{\emptyset\}}$ of a game $(N,v)$, that is:

$$\phi_i(v)=\sum_{S\subseteq N:i\in S}\frac{\lambda_S(v)}{|S|} \tag{2}$$

for each $i\in N$.

### 2.2 Genes and games

Suppose to have a set $K$ of *key genes* assumed to be *equally important* for the regulation of a certain biological process. Let $N$ be the set of genes who are studied together with genes in $K$ on a sequence of (microarray) experiments under a condition of interest, for instance, a genetic disorder. Let $I \subseteq \{\{i,k\} | i \in N, k \in K\}$ be the set of *interactions* between genes in $N$ and key genes in $K$. We will say that a gene $i \in N$ and a key gene $k \in K$ *interact* if and only if $\{i,k\} \in I$. The triple $(N,K,I)$ is said a *gene-k-gene* (gkg) situation.

Given a set of genes $S \subseteq N$, the higher the number of key genes which interact with genes in $S$, the higher the likelihood that genes in $S$ are also involved in the regulation of the biological process of interest. In order to measure the strength of *association* of pathways of genes in $N$, for each group $S \subseteq N$ we compute the number of key genes interacting only with genes in $S$. Let $v : 2^N \to \mathbb{N}$ be the map assigning to each coalition $S \in 2^N \setminus \{\emptyset\}$ the number $v(S)$ of key genes in $K$ which only interact (in $I$) with genes in $S$. By convention, $v(\emptyset) = 0$. The pair $(N,v)$ is called *association game* corresponding to $(N,K,I)$. Note that the assumption of equal importance for key genes is central for the definition of the characteristic function $v$. In fact, the value $v(S)$, for each $S \in 2^N \setminus \{\emptyset\}$, represents a measure of the relevance of coalition $S$ in terms of the number of key genes directly interacting only with genes in $S$. The possibility to compare the relevance of different coalitions makes sense thanks to the assumption of equal importance of key genes.

In the remainder of the article, to simplify the presentation of the game-theoretic model, we will also assume that key genes are *independent*, i.e. they do not directly interact between them. However, this assumption is not fundamental as the one of equal importance. If a group of $m$ key genes directly interact, it will be sufficient to collapse them into an individual key-unit whose importance equals $m$ times the importance of a single key genes.

Example 1. *Consider a set of genes* $N = \{1,2,3,4\}$, *a set of key-genes* $K = \{a,b,c\}$ *and a set of interactions* $I = \{\{1,a\},\{1,b\},\{3,b\},\{3,c\}, \{4,c\}\}$, *as they are depicted by thin lines in Figure 1.*

*The association game* $(N,v)$ *is such that* $v(\emptyset) = v(2) = v(3) = v(4) = v(2,3) = v(2,4) = 0$, $v(1,3) = v(1,2,3) = 2$, $v(1,3,4) = v(1,2,3,4) = 3$ *and* $v(S) = 1$ *for all the remaining coalitions.*

If gene $i \in N$ has not directly an interaction with $k \in K$, it may still be possible for $i$ to interact with $k$ via an interaction with another gene $j \in N$ (an intermediary) which in turn has an interaction with $k$, or more generally, via a sequence of intermediaries. So, it is essential to understand which genes really interact, directly or via intermediaries, and how the network of such interactions may affect the worth of coalitions of genes.

Let us consider now an *interaction network* $\langle N, \Gamma \rangle$, the nodes of the graph being the genes. The set of edges $\Gamma$ indicates interaction ties between pairs of genes, i.e. a set $\{i,j\} \subseteq N$ is an element of $\Gamma$ if and only if $i$ and $j$ have an interaction. Implicitly, this graph shows us which coalitions are feasible, i.e. which coalitions have all their members related by interactions.

Given a gkg situation $(N,K,I)$ with the corresponding association game $(N,v)$ and an interaction network $\langle N, \Gamma \rangle$, following the approach in (Myerson, 1977), we use the structure of an interaction network to define a new game $(N, w^v_\Gamma)$, where the value $w^v_\Gamma(S)$ of a coalition $S$ equals the sum of the values assigned by $v$ to the connected components of the network restricted to this coalition $S$. The game $w^v_\Gamma$ is called the graph-restricted game.

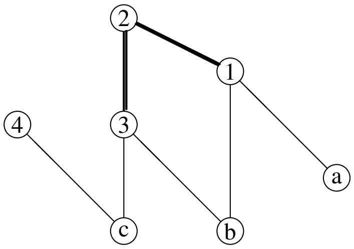


**Fig. 1.** Interaction network $\hat{\Gamma}$ (thick lines) and the interactions of the gkg situation described in Example 1 (thin lines).

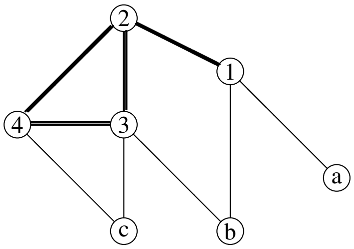


**Fig. 2.** Interaction network $\bar{\Gamma}$ (thick lines) and the interactions of the gkg situation described in Example 1 (thin lines).

Definition 1. *Let* $(N,K,I)$ *be a gkg situation and let* $(N,v)$ *be the corresponding association game. Let* $\langle N, \Gamma \rangle$ *be an interaction network. The* graph-restricted game $(N, w^v_\Gamma)$ *is defined by*

$$w^v_\Gamma(S) = \sum_{T \in C_{\Gamma_S}} v(T) \tag{3}$$

*for each* $S \in 2^N \setminus \{\emptyset\}$, *where* $C_{\Gamma_S}$ *is the set of all the connected components in* $\langle S, \Gamma_S \rangle$, *and with the convention* $w^v_\Gamma(\emptyset) = 0$.

Example 2. *Consider the gkg situation of Example 1 with the corresponding association game* $(N,v)$. *Consider the interaction network* $(N, \hat{\Gamma})$ *where* $\hat{\Gamma} = \{\{1,2\},\{2,3\}\}$. *All the interactions are represented in network of Figure 1.*

*The graph-restricted game* $(N, w^v_{\hat{\Gamma}})$ *is such that* $w^v_{\hat{\Gamma}}(1) = w^v_{\hat{\Gamma}}(1,2) = w^v_{\hat{\Gamma}}(1,4) = w^v_{\hat{\Gamma}}(1,2,4) = w^v_{\hat{\Gamma}}(1,3) = w^v_{\hat{\Gamma}}(1,3,4) = 1$, $w^v_{\hat{\Gamma}}(1,2,3) = 2$, $w^v_{\hat{\Gamma}}(1,2,3,4) = 2$ *and* $w^v_{\hat{\Gamma}}(S) = 0$ *for all the remaining coalitions.*

Example 3. *Consider the gkg situation of Example 1 with the corresponding association game* $(N,v)$. *Consider the interaction network* $(N, \bar{\Gamma})$ *where* $\bar{\Gamma} = \{\{1,2\},\{2,3\},\{2,4\},\{3,4\}\}$. *All the interactions are represented in network of Figure 2.*

*The graph-restricted game* $(N, w^v_{\bar{\Gamma}})$ *is such that* $w^v_{\bar{\Gamma}}(3,4) = w^v_{\bar{\Gamma}}(2,3,4) = 1$, $w^v_{\bar{\Gamma}}(1) = w^v_{\bar{\Gamma}}(1,2) = w^v_{\bar{\Gamma}}(1,4) = w^v_{\bar{\Gamma}}(1,2,4) = w^v_{\bar{\Gamma}}(1,3) = 1$, $w^v_{\bar{\Gamma}}(1,2,3) = w^v_{\bar{\Gamma}}(1,3,4) = 2$, $w^v_{\bar{\Gamma}}(1,2,3,4) = 3$ *and* $w^v_{\bar{\Gamma}}(S) = 0$ *for all the remaining coalitions.*

Starting from the basic paper (Shapley and Shubik, 1954), the Shapley value of a game has been considered as a player's power in several different applications [see, for instance, the survey (Moretti and Patrone, 2008) for references to the use of the Shapley value as a power index in different contexts]. Here, players are genes and the Shapley value is considered as a gene's power. The intuition behind the meaning of gene's power attributed to relation (1) follows from this consideration. An order $\sigma$ on a set of genes $N$ may be interpreted as a sequence of activations of study genes and the corresponding marginal vector may be seen as a measure of the power of study genes to establish relevant interactions with key genes according to $\sigma$. However, in absence of information about which sequences of activations are more likely, it is reasonable to average the marginal vectors over all possible orders as an indication of the expected power of genes.

The difference between the power of a gene in the graph-restricted game and its power in the association one is proposed as a centrality measure for co-expression networks [see (Gòmez *et al.*, 2003) in the context of social networks]. Let $(N,K,I)$ be a gkg situation and let $(N,v)$ be the corresponding association game. Let $\langle N,\Gamma\rangle$ be an interaction network. The *centrality measure* $\gamma(v,\Gamma)$ is defined by

$$\gamma_i(v,\Gamma)=\phi_i(w^v_\Gamma)-\phi_i(v), \qquad (4)$$

for each $i\in N$, where $\phi(v)$ is the Shapley value of the association game $v$ and $\phi(w^v)$ is the Shapley value of the corresponding graph-restricted game $w^v_\Gamma$. According to relation (4), genes with strictly positive $\gamma$ represent those genes with a positive differential power between the graph-restricted game and the association game. In the applications introduced in Sections 3 and 4, we will be interested to study such genes, because they are genes whose power increases as a consequence of their interactions in the network.

Example 4. *Consider the gkg situation with the corresponding association game $(N,v)$ and the interaction network of Example 2. According to relation (1), we have that $\phi(v)=(\frac{3}{2},0,1,\frac{1}{2})$ and $\phi(w^v_{\hat\Gamma})=(\frac{4}{3},\frac{1}{3},\frac{1}{3},0)$. Thus, the centrality measure gives $\gamma(v,\hat\Gamma)=(-\frac{1}{6},\frac{1}{3},-\frac{2}{3},-\frac{1}{2})$.*

*As an example of Shapley value computation via relation (1), we show here the calculation for gene* 1. *In total, there are* $4!=24$ *orders in $\Sigma_N$. There are precisely* 6 *orders $\sigma\in\Sigma_N$ such that $\sigma^{-1}(1)=1$ and other* 6 *orders $\sigma\in\Sigma_N$ such that $\sigma^{-1}(4)=1$. In addition, for each intermediate coalition $S\subseteq\{2,3,4\}$ of one or two genes, there are two orders on N such that S is the set of precessors of* 1. *Consequently, from relation (1), the Shapley value of gene* 1 *is*

$$\begin{aligned}\phi_1(v)=&\tfrac{1}{24}\big(6(v(\{1)-v(\emptyset))+6(v(1,2,3,4)-v(2,3,4))\\&+2(v(1,2)-v(2))+2(v(1,3)-v(3))+2(v(1,4)-v(4))\\&+2(v(1,2,3)-v(2,3))+2(v(1,3,4)-v(3,4))\\&+2(v(1,2,4)-v(2,4))\big)\\=&\tfrac{1}{24}\big(6\times(1-0)+6\times(3-1)+2\times(1-0)+2\times(2-0)\\&+2\times(1-0)+2\times(2-0)+2\times(3-1)+2\times(1-0)\big)\\=&\tfrac{1}{24}\big(6+12+2+4+2+4+4+2\big)=\tfrac{36}{24}=\tfrac{3}{2}.\end{aligned}$$

Next section is devoted to illustrate a more efficient way to calculate the Shapley value of genes.

Example 5. *Consider the gkg situation with the corresponding association game $(N,v)$ and the interaction network of Example 3. Again, according to relation (1), we have that $\phi(v)=(\frac{3}{2},0,1,\frac{1}{2})$ (nothing changed in game v) and $\phi(w^v_{\bar\Gamma})=(\frac{4}{3},\frac{1}{3},\frac{5}{6},\frac{1}{2})$. Thus, the centrality measure gives $\gamma(v,\bar\Gamma)=(-\frac{1}{6},\frac{1}{3},-\frac{1}{6},0)$. Note that with respect to Example 4, where edges* $\{2,4\}$ *and* $\{3,4\}$ *were not present, gene* 2 *continues to be the unique one with strictly positive centrality according to $\gamma$, even if genes* 3 *and* 4 *increase their centrality.*

It should be noted that $-v(N)\leq\gamma_i(v,\Gamma)\leq w^v_\Gamma(N)$ for each $i\in N$. As a consequence, $\gamma$ centrality computed on different interaction networks are comparable scores only if they are defined on the same interval scale, that is if the worth of the largest coalition in the graph-restricted game is the same for both interaction networks.

## 2.3 Centrality computation

Actually, the computation of the Shapley value using relation (1) may be very hard even if the number of genes is quite small. For instance, with only 10 genes, relation (1) needs $10!=3628800$ orders of genes. In order to make real applications, it is useful to decompose the association game and the corresponding graph-restricted game by means of a relatively small number of unanimity games with non-null unanimity coefficients. As a consequence, the Shapley value of such games may be computed in a less complex way via relation (2). In the following, we briefly describe this decomposition procedure.

Let $(N,K,I)$ a gkg situation. For each key gene $k\in K$, the set of genes in $N$ which have a strong interaction with $k$ are denoted by $N_k=\{i\in N|\{i,k\}\in I\}$. In the remainder of the article, genes in $N_k$, for each $k\in K$, will be shortly referred as *most associated genes*. Let $(N,v)$ be the corresponding association game. It is easy to show that the characteristic function $v$ can be written as a sum of unanimity games:

$$v=\sum_{k\in K,N_k\neq\emptyset}u_{N_k}. \qquad (5)$$

Example 6. *Consider the gkg of Example 1. We have that* $N_a=\{1\},N_b=\{1,3\},N_c=\{3,4\}$. *From relation (5), the corresponding association game v is given by*

$$v=u_{\{1\}}+u_{\{1,3\}}+u_{\{3,4\}}.$$

*Consequently, according to relation (2) (with unanimity coefficients $\lambda_S(v)=1$, if $S\in\{\{1\},\{1,3\},\{\{3,4\}\}$, and $\lambda_S(v)=0$, otherwise) , the Shapley value of v can easily be calculated as the following sum of vectors*

$$\phi(v)=(1,0,0,0)+(\tfrac{1}{2},0,\tfrac{1}{2},0)+(0,0,\tfrac{1}{2},\tfrac{1}{2})=(\tfrac{3}{2},0,1,\tfrac{1}{2}).$$

The remainder of this section is devoted to provide a natural decomposition of a graph-restricted game based on the reformulation of the association game given in (5).

First, we need to introduce the concept of minimal component containing a coalition $S$. Let $\langle N,E\rangle$ be a graph. We denote by $\langle N,F_E\rangle$ a graph where $F_E$ is a maximal subset of $E$ with the property that $\langle N,F_E\rangle$ is a forest. The set of all the forests for $E$ is denoted by $\mathcal{F}_E$. Let $S\in 2^N\setminus\{\emptyset\}$. A *minimal component containing* $S$ in a forest $\langle N,E\rangle$ is a minimal subset of $N$ which contains $S$ and with the property that any two nodes in this set are connected in $\langle N,E\rangle$. Note that in a forest, a minimal component containing $S$, if exists, is unique. This fact allows us to denote the minimal component containing $S$ in a forest $F_E$ (if it exists) by $M_{F_E}(S)$, and the set of all the minimal components containing $S$ in a graph $\langle N,E\rangle$ is denoted by $\mathcal{M}_E(S)=\{M_{F_E}(S)|F_E\in\mathcal{F}_E\}$.

Let $\langle N,\Gamma\rangle$ be a graph. Consider a unanimity game $(N,u_S)$, with $S\in 2^N\setminus\{\emptyset\}$ and such that $\mathcal{M}_\Gamma(S)\neq\emptyset$. Without loss of generality, suppose that $\mathcal{M}_\Gamma(S)=\{M^{i_1}_\Gamma(S),\ldots,M^{i_r}_\Gamma(S)\}$, with $r\geq 1$. We define a new game $(N,w^{u_S}_\Gamma)$ in the following way

$$w^{u_S}_\Gamma=\sum_{j=1}^{r}(-1)^{j+1}\sum_{1\leq i_1<\cdots<i_j\leq r}u_{M^{i_1}_\Gamma(S)\cup\cdots\cup M^{i_j}_\Gamma(S)}. \qquad (6)$$

Example 7. *Consider the interaction network of Example 3. Let* $S=\{1,3\}$. *Note that* $\mathcal{M}_{\bar\Gamma}(S)=\{\{1,2,3\},\{1,2,3,4\}\}$. *From relation (6) we have that*

$$w_{\Gamma}^{u_S}=u_{\{1,2,3\}}+u_{\{1,2,3,4\}}-u_{\{1,2,3\}\cup\{1,2,3,4\}}=u_{\{1,2,3\}}.$$

Games defined according to relation (6) are crucial for the computation of the Shapley value of graph-restricted games in practical situations. In fact, it can be proved that the game $w_{\Gamma}^{u_S}$ is the restriction of the unanimity game $u_S$ to graph $\Gamma$, and is also known as the connecting $S$ in $\Gamma$ game (Gòmez *et al*., 2004). Consequently, it can be easily shown that given a gkg situation $(N,K,I)$ with the corresponding association game $(N,v)$ and an interaction network $\langle N,\Gamma\rangle$, the graph-restricted game $(N,w_{\Gamma}^{v})$ may be computed via the following formula

$$\begin{aligned}
&w_{\Gamma}^{v}=\\
&=\textstyle\sum_{k\in K,N_k\neq\emptyset,\mathcal{M}_{\Gamma}(N_k)\neq\emptyset} w_{\Gamma}^{u_{N_k}}\\
&=\textstyle\sum_{k\in K,N_k\neq\emptyset,\mathcal{M}_{\Gamma}(N_k)\neq\emptyset}\\
&\textstyle\sum_{j=1}^{|\mathcal{M}_{\Gamma}(N_k)|}(-1)^{j+1}\sum_{1\leq i_1<\cdots<i_j\leq r} u_{M_{\Gamma}^{i_1}(N_k)\cup\cdots\cup M_{\Gamma}^{i_j}(N_k)},
\end{aligned}\tag{7}$$

where $\mathcal{M}_{\Gamma}(N_k)=\{M_{\Gamma}^{i_1}(N_k),\ldots,M_{\Gamma}^{|\mathcal{M}_{\Gamma}(N_k)|}(N_k)\}$, for each $k\in K$ with $N_k\neq\emptyset$ and $\mathcal{M}_{\Gamma}(N_k)\neq\emptyset$.

From relations (2) and (7), it immediately follows that the Shapley value of a graph-restricted game $w_{\Gamma}^{v}$ can be computed using the following relation

$$\begin{aligned}
&\phi_i(w_{\Gamma}^{v})\\
&=\textstyle\sum_{k\in K,i\in N_k,\mathcal{M}_{\Gamma}(N_k)\neq\emptyset}\\
&\textstyle\sum_{j=1}^{|\mathcal{M}_{\Gamma}(N_k)|}(-1)^{j+1}\sum_{1\leq i_1<\cdots<i_j\leq r}\frac{1}{|M_{\Gamma}^{i_1}(N_k)\cup\cdots\cup M_{\Gamma}^{i_j}(N_k)|},
\end{aligned}\tag{8}$$

for each $i\in N$.

Example 8. *Consider the gkg with the corresponding association game* $(N,v)$ *and the interaction network of Example 2. Note that* $\mathcal{M}_{\hat\Gamma}(N_a)=\{\{1\}\}$, $\mathcal{M}_{\hat\Gamma}(N_b)=\{\{1,2,3\}\}$ *and* $\mathcal{M}_{\hat\Gamma}(N_c)=\{\emptyset\}$.

*According to relation (7), we can write the graph-restricted game* $w_{\hat\Gamma}^{v}$ *as a sum of unanimity games*

$$w_{\hat\Gamma}^{v}=u_{\{1\}}+u_{\{1,2,3\}}.\tag{9}$$

*Consequently,* $\phi(w_{\hat\Gamma}^{v})=(\frac{4}{3},\frac{1}{3},\frac{1}{3},0)$.

Example 9. *Consider the gkg with the corresponding association game* $(N,v)$ *and the interaction network of Example 3. Note that* $\mathcal{M}_{\Gamma}(N_a)=\{\{1\}\}$, $\mathcal{M}_{\Gamma}(N_b)=\{\{1,2,3\},\{1,2,3,4\}\}$ *and* $\mathcal{M}_{\Gamma}(N_c)=\{\{3,4\},\{2,3,4\}\}$.

*According to relation (7), we can write the graph-restricted game* $w_{\Gamma}^{v}$ *as a sum of unanimity games*

$$\begin{aligned}
&w_{\Gamma}^{v}=u_{\{1\}}\\
&+u_{\{1,2,3\}}+u_{\{1,2,3,4\}}-u_{\{1,2,3\}\cup\{1,2,3,4\}}\\
&+u_{\{3,4\}}+u_{\{2,3,4\}}-u_{\{3,4\}\cup\{2,3,4\}}\\
&=u_{\{1\}}+u_{\{1,2,3\}}+u_{\{3,4\}}.
\end{aligned}\tag{10}$$

*Consequently,* $\phi(w_{\Gamma}^{v})=(\frac{8}{6},\frac{2}{6},\frac{5}{6},\frac{3}{6})$. *Note that the role of the minimal components* $\{1,2,3,4\}$ *and* $\{2,3,4\}$ *in the computation of the graph-restricted game* $w_{\Gamma}^{v}$ *is* redundant *(both of them are superset of 'smaller' minimal component, and consequently the contribution of the corresponding unanimity games to the computation of* $w_{\Gamma}^{v}$ *is null). Such redundant components may be* a priori *eliminated from the analysis, with a consequent reduction of computational burden in relations (7) and (8).*

In the next section, we present an application of this centrality measure on a microarray data from children exposed to air pollution (Moretti *et al*., 2008).

## 3 PRELIMINARY APPLICATION

We present a preliminary application of the method to gene expression data published in (van Leeuwen *et al*., 2008), where genome-wide oligonucleotide microarray analysis was applied to blood cells of 23 children from Teplice (TP) region in the Czech Republic. The TP region is a mining district characterized by high levels of airborne pollutants including carcinogens. We consider the gene expression matrix **X** of 20130 genes and 23 samples from TP that was distilled from the data filtering and preparation as described in (van Leeuwen *et al*., 2008).

As a set of key genes, we used four genes known to be strongly associated with micronuclei frequencies, a biomarker of chromosome damage: (i) PRC1 (protein regulator of cytokinesis 1); (ii) TP53 [tumor protein p53 (li-fraumeni syndrome)]; (iii) ZWINT (zw10 interactor); and (iv) CCNB2 (cyclin b2). As a first filtering step, absolute values of Pearson's correlation coefficients between each study gene and each key gene were computed, providing four lists of correlation coefficients (one list for each key gene) with 20130 genes each, and the union of the top 25 genes from the four lists were selected for further analysis ($n=96$). From the gene expressions of the selected 96 genes, the corresponding gene correlation matrix was computed, and an unweighted network, based on dichotomizing the correlation matrix, was considered. More precisely, two genes were considered to interact (i.e. linked by an edge in the network) if and only if their absolute Pearson's correlation coefficient was >0.75 (Fig. 3).

According to this criterion, it was possible to define the association game on the total set of 96 genes as the set of players, and the corresponding graph-restricted game. From the association game, only 9 genes obtained a non-null Shapley value ranging from 1 to 0.25 (Fig. 3, most associated genes). In fact, from relation (5), the association game is defined by

$$v=u_{\{1,2,3,4\}}+u_{\{26,27,28\}}+u_{\{38\}}+u_{\{72\}}$$

and, as a consequence of relation (2), $\phi_1(v)=\phi_2(v)=\phi_3(v)=\phi_4(v)=0.25$, $\phi_{26}(v)=\phi_{27}(v)=\phi_{28}(v)=\frac{1}{3}$, $\phi_{38}(v)=\phi_{72}(v)=1$ and $\phi_i(v)=0$ for each other gene $i$.

In order to compute the Shapley value on the graph-restricted game, as it was described in Section 2.3, we first should find the sets of minimal connected components $\mathcal{M}_{\Gamma}(N_k)$, for each key gene $k$. By definition, this requires the computation of the set of all the forests $\mathcal{F}(\Gamma)$. Several algorithms exist for generating all spanning trees of a graph that can be easily adapted to find all the forests [e.g. (Gabow and Myers, 1978; Kapoor and Ramesh, 1995; Minty, 1965). However, as the number of forests in a graph can be very large (especially for graphs generated from datasets with thousands of genes) this option is excluded for practical purposes on large datasets.

In this preliminary application, the computation of the Shapley value on the graph-restricted game may be done by means of visual inspection of the graph, looking at the shortest paths (thicker edges

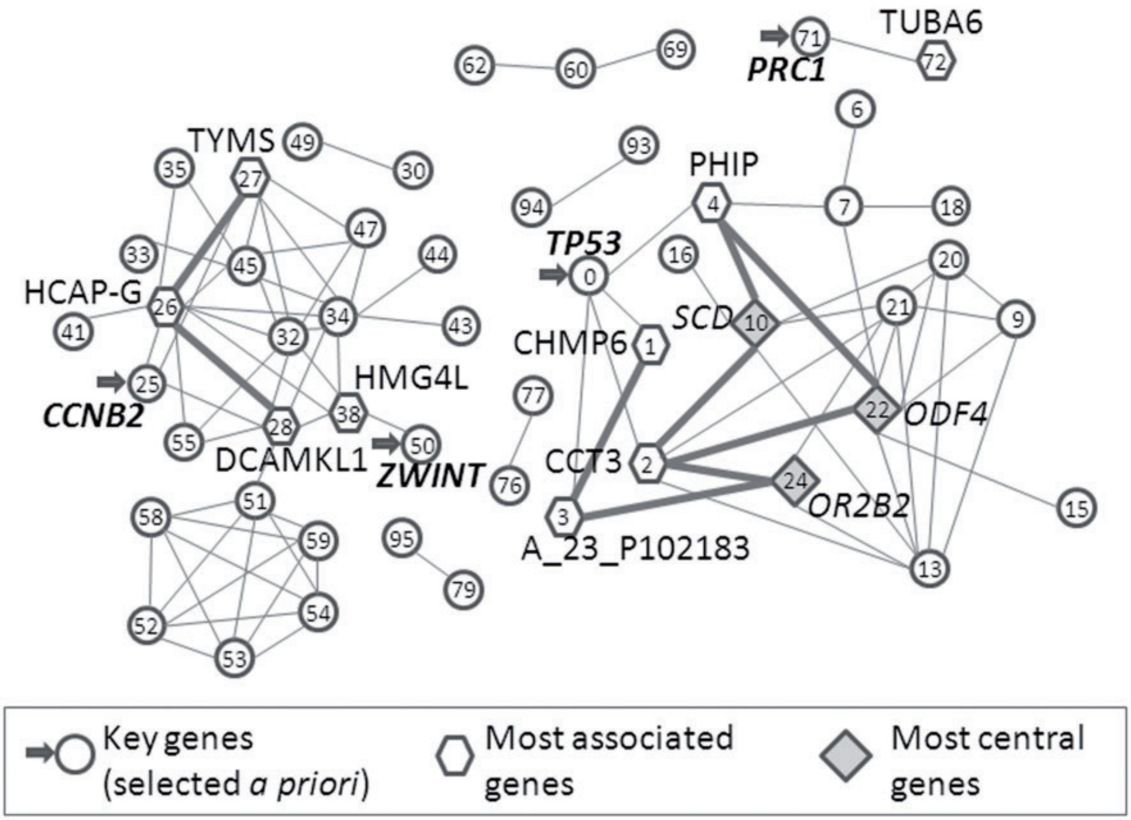


**Fig. 3.** Interaction network between genes (nodes). Interactions between gene's pairs are represented by edges. Isolated genes were removed. Thicker edges show the shortest paths among the most associated genes. Most central genes according to the $\gamma$ measure of centrality are shown.

in Fig. 3), which connect nodes of unanimity coalitions (the most associated genes in Fig. 3). It is in fact easy to check that all minimal components containing the unanimity coalition $\{1,2,3,4\}$ must include the component $\{1,2,3,4,10,24\}$ or the component $\{1,2,3,4,22,24\}$. Consequently, by relation (7), the graph-restricted game is

$$w_\Gamma^v = u_{\{1,2,3,4,10,24\}} + u_{\{1,2,3,4,22,24\}} - u_{\{1,2,3,4,10,22,24\}} \\ + u_{\{26,27,28\}} + u_{\{38\}} + u_{\{72\}}$$

and, by relation (2), $\phi_1(w_\Gamma^v) = \phi_2(w_\Gamma^v) = \phi_3(w_\Gamma^v) = \phi_4(w_\Gamma^v) = \phi_{24}(w_\Gamma^v) = \frac{2}{6} - \frac{1}{7} = \frac{4}{21}$, $\phi_{10}(w_\Gamma^v) = \phi_{22}(w_\Gamma^v) = \frac{1}{6} - \frac{1}{7} = \frac{1}{42}$, $\phi_{26}(w_\Gamma^v) = \phi_{27}(w_\Gamma^v) = \phi_{28}(w_\Gamma^v) = \frac{1}{3}$, $\phi_{38}(w_\Gamma^v) = \phi_{72}(w_\Gamma^v) = 1$ and $\phi_i(w_\Gamma^v) = 0$ for each other gene $i$.

By relation (4) and the above calculations, only three genes have a $\gamma$ centrality measure larger than zero, i.e. OR2B2, SCD and ODF4 (Table 1 and Fig. 3, most central genes). The same value of $\gamma$ for SCD and ODF4 can be explained by the fact that these two genes can be alternatively used to connect genes PHIP and CCT3 (Fig. 3, thicker edges). Instead, OR2B2 plays a more critical role, since it is necessary to connect genes A_23_P102183 and CCT3.

Such genes are connected to genes associated to the key gene TP53. This is a consequence of the fact the other three key genes do not contribute to $\gamma$ centrality, being the terms $\phi(u_{\{26,27,28\}}) + \phi(u_{\{38\}}) + \phi(u_{\{72\}})$ both in $\phi(v)$ and $\phi(w_\Gamma^v)$.

Among genes with positive $\gamma$, gene OR2B2 encodes for an olfactory receptor protein which is member of a large family of G-protein-coupled receptors. G proteins have been suggested to be involved in the respiratory burst (release of ROS) caused by asbestos (Elferink and Ebbenhout, 1988). In addition, it is known that fine and ultra-fine particulate matter air pollution may reach the brain through olfactory receptor neurons and the trigeminal nerves (Calderòn-Garcidueña *et al.*, 2007). The principal product of SCD is oleic acid, which is formed by desaturation of stearic acid. The ratio of stearic acid to oleic acid has been implicated in the regulation of cell growth and differentiation through effects on cell membrane fluidity and signal transduction. ODF4 encodes a protein that is localized in the outer dense fibers of the tails of mature sperm. As a functional annotation, all such genes encode for transmembrane proteins.

**Table 1.** Genes with $\gamma$ centrality measure greater than zero

| ID | Symbol | Name | $\gamma$ Centrality |
|---|---|---|---|
| 24 | OR2B2 | Olfactory receptor, family 2, subfamily B, member 2 | $\frac{4}{21}$ |
| 10 | SCD | Stearoyl-CoA desaturase (delta-9-desaturase) | $\frac{1}{42}$ |
| 22 | ODF4 | Outer dense fiber of sperm tails 4 | $\frac{1}{42}$ |

Methods were implemented using R language (R Development Core Team, 2004).

## 4 APPROXIMATE COMPUTATION

As we already observed in the last section, the implementation of algorithms aimed to generate the set of all forests in a real biological network is an unpractical approach because of the huge storage memory and the computational burden. This section is devoted to the description of an alternative approach based on approximate calculations, and to its application on a large biological network where also other centrality measures are applied.

Let $(N,K,I)$ be a gkg situation and let $\langle N,\Gamma\rangle$ be an interaction network. For instance, with $|N|=15$, $|K|=1$, $|I|=5$ and $|\Gamma|=21$ (i.e. $\langle N,\Gamma\rangle$ has a graph density equal to 0.2), the exact computation of the Shapley value of the restricted game $w_\Gamma^v$ according to relation (8) and our R language implementation, required <4 min (on a PC with a 2 GHz processor and 2 GB of memory). But the problem explodes exponentially in time on more dense graphs. Even if the time required for the computation is specific to the implementation, the main complexity issue deals with the number of operations needed to find all forests. Note that for a complete graph the Cayley's formula states that there are precisely $n^{n-2}$ spanning trees, where $n$ is the number of nodes; in general, if a connected graph is not complete, the number of spanning trees depends on the structure of the graph, and it can be computed using the Kirchhoff's matrix-tree theorem (see, for instance, Bondy and Murty, 1976). For certain classes of graphs, such a number may be bounded from above: for instance, a $k$-regular connected graph on $n$ vertices contains less than $k^n$ spanning trees (Alon, 1990) or, alternatively, less than $(n-1)^n d^n$ (where $d$ is the graph density of a $k$-regular connected graph on $n$ vertices).

For this reason, in order to make feasible (and reasonable in terms of elapsed time) the application of the method also to larger biological networks, we avoided the exhaustive generation of all forests and the consequent exact computation of the Shapley value of a restricted game. Alternatively, we limited the computation of relation (8) to the 'smallest' minimal components connecting the most associated genes on a graph, i.e. to those components which belong to forests where each most associated gene is connected to another most associated one by a shortest path.

DEFINITION 2. *Let $\langle N,\Gamma\rangle$ be an interaction network. For each forest $F_\Gamma\in\mathcal{F}(\Gamma)$ and each pair of nodes $i,j\in N, i\neq j$, let $F_\Gamma^{i,j}$ be the unique path in $F_\Gamma$ between i and j. For each $S\in 2^N\setminus\{\emptyset\}$,the set of* smallest minimal components *is defined as the set*

$$\begin{aligned}\bar{\mathcal{M}}_\Gamma(S)=\{M_{F_\Gamma}(S)|F_\Gamma\in\mathcal{F}(\Gamma)\text{ and}\\ \forall\, i\in S\ \exists\, j\in S, i\neq j,\ \text{s.t. } F_\Gamma^{i,j}\in\mathcal{SP}_E^{i,j}\}.\end{aligned}\tag{11}$$

EXAMPLE 10. *Consider again the interaction situation depicted in Figure 3. If we consider the set of smallest minimal components for coalition $N_0=\{1,2,3,4\}$ (i.e. the genes most associated to TP53), we have that $\bar{\mathcal{M}}_\Gamma(N_0)=\{\{1,2,3,4,10,24\},\{1,2,3,4,22,24\}\}$. Instead, if we consider the set of most associated genes $N_{71}=\{72\}$ (i.e. TUBA6 which is the unique gene most associated to PRC1), we have that $\bar{\mathcal{M}}_\Gamma(N_{71})=\{\{72\}\}$. In fact, the second condition in relation (11) is always satisfied if $|S|=1$ (there are no distinct elements in S), and therefore $\bar{\mathcal{M}}_\Gamma(N_{71})=\mathcal{M}_\Gamma(N_{71})$ (a similar reasoning may be done for gene HMG4L in the role of PRC1).*

*To complete the example, note that $\bar{\mathcal{M}}_\Gamma(N_{71})=\{\{26,27,28\}\}$.*

The *approximate* Shapley value of $w_\Gamma^v$ is computed according to Equation (8) with $\bar{\mathcal{M}}_\Gamma$ in the role of $\mathcal{M}_\Gamma$.

This procedure was used to calculate an approximate $\gamma$ centrality for a larger graph with 201 nodes and 2083 edges. Only one key gene (again gene TP53) was considered, on the same dataset introduced in Section 3. In this case, 250 genes with the highest absolute value of Pearson's correlation with TP53 were initially selected for further analysis. From the gene expressions of the selected 250 genes, following the same method described in Section 3, a network was constructed. More precisely, a link between two nodes was established if and only if their absolute Pearson's correlation coefficient was greater than 0.75. Only genes connected (directly or via other nodes) to TP53 were considered (finally, $|N|=201$). We focused exclusively on the component connected to key gene TP53 because the contribution of the other key genes to $\gamma$ centrality in a larger network (constructed according to the procedure previously described) is the same as it was calculated at the end of Section 3 on the network depicted in Figure 3, that is null.

The algorithm for the approximate computation of $\gamma$ centrality was applied to the generated interaction network (elapsed time 29.3 min).

The number of smallest minimal components connecting the most associated genes A_23_P102183, CHMP6, CCT3 and PHIP, was 105. This number was calculated combining in all possible ways all the shortest paths between each pair of most associated genes, as obtained by the application of the R function *get.all.shortest.paths()* in the package *igraph* (Csardi and Nepusz, 2006), under the condition that the resulting graphs were forests. Many of these smallest minimal components were redundant (Example 9) and therefore removed before the computation of relation (8). After removal, the number of non-redundant components used for computation in relation (8) was 21. This is in fact the upper bound of $|\mathcal{M}_\Gamma|$ for the exact computation of the Shapley value according to our R implementation of relation (8).

Only nine genes showed an approximate $\gamma$ centrality strictly positive (indeed, representing genes with a positive differential power). Those findings were compared with the most nine central genes according to other four common measures of centrality. In the following, we briefly introduce those measures. In order to do that, we denote by $d(u,v)$ the minimum number of edges to connect two nodes $u$ and $v$ in $\langle N,\Gamma\rangle$:

(1) Degree centrality (Nieminen, 1974; Shaw, 1954): the degree centrality of $v\in N$ is defined as the number of edges in $e\in\Gamma$ such that $v\in e$.

(2) Closeness centrality (Beauchamp, 1965; Sabidussi, 1966): the closeness centrality of node $v\in N$ is defined as $\frac{|N|-1}{\sum_{y\in N}\mathrm{d}(v,y)}$. Therefore, it measures the extent to which node $v\in N$ is close to all other nodes in the $\langle N,\Gamma\rangle$.

(3) Betweenness centrality (Bavelas, 1948; Freeman, 1977): let $u,v,z\in N$ and let $n_{u,v}$ be the number of paths formed by precisely $\mathrm{d}(u,v)$ edges and let $n_{u,v}(z)$ be the number of paths formed by precisely $\mathrm{d}(u,v)$ edges which contains node $z$. The rate of communication between $u$ and $v$ that can be monitored by an interior node $z$ is denoted by $\delta_{u,v}(z)=n_{u,v}(z)/n_{u,v}$. If no shortest path between $u$ and $v$ exists $\delta_{u,v}(z)=0$ by definition. The betweenness centrality of $z$ is defined as $\sum_{u,v\in N,u\neq v,u\neq z,v\neq z}\delta_{u,v}(z)$.

(4) Eigenvector centrality (Bonacich, 1972): let $v\in N$. Then the eigenvector centrality of $v$ is defined as the $v$-th element of the

**Table 2.** Common findings among lists of nine genes with highest centrality according to different centrality measures

| | Appr. $\gamma$ | Deg. | Clos. | Bet. | Eigen. |
|---|---|---|---|---|---|
| approxim. $\gamma$ | 9 | 2 | 4 | 6 | 0 |
| Deg. | 2 | 9 | 7 | 3 | 7 |
| Clos. | 4 | 7 | 9 | 5 | 5 |
| Bet. | 6 | 3 | 5 | 9 | 1 |
| Eigen. | 0 | 7 | 5 | 1 | 9 |

principal eigenvector of the adjacency matrix corresponding to $\langle N, \Gamma \rangle$. This principal eigenvector is normalized such that its largest entry is 1. This centrality is a measure for how well connected a node is to other highly connected nodes in a network.

The time needed to compute the nodes' centrality according to each of the four measures described above was $<1$ s, using the R functions in the package *igraph* (Csardi and Nepusz, 2006).

For each pair of centrality measures considered, the number of common genes among the first nine with highest centrality for each measure are reported in Table 2. Note that the betweenness centrality has the maximum level of overlap with the list of genes ranked according to the approximate $\gamma$ centrality.

The nine central genes selected according to the approximate $\gamma$ are reported on Table 3. Note the presence of genes SCD, ODF4 and OR2B2, that were found in the analysis of the smaller network introduced in Section 3 (see Table 3, where indications from other centrality measures are shown too). Among genes that are predicted to be central only by gamma centrality (namely, SFPQ, OR2B2 and THRAP1), we observe that SFPQ is a multifunctional protein that has been suggested to play a role in tumorigenesis, as SFPQ translocation occurs in papillary renal cell carcinoma (Clark *et al.*, 1997; Rubin and Sive, 2007). Moreover, nuclear relocalization and hyperphosphorylation of the protein that SFPQ encodes [known as polypyrimidine tract binding protein-associated splicing factor (PSF)] occur during apoptosis (Shav-Tal *et al.*, 2001). PSF has been shown to exhibit multiple functions in nucleic acid synthesis and processing *in vitro* and in tissue culture, including RNA polymerase II. Similarly, THRAP1 (also known as MED13) is a component of the Mediator complex, a co-activator involved in the regulated transcription of nearly all RNA polymerase II-dependent genes. The role of gene OR2B2 was discussed at the end of the previous section.

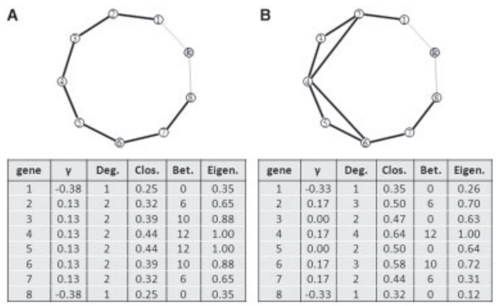
From Tables 2 and 3, it seems that $\gamma$ centrality behaves very close to betweenness centrality, at least with respect to the most central genes (six genes in common among the top nine).

This result is not surprising, but it can be explained by means of the very basic properties of the $\gamma$ index. The definition of $\gamma$ centrality, based on the notions of minimal components for coalitions (see relation 8), gives more importance to geodesic paths which connect the most associated genes. This is in fact a generalization of the notion of betweenness centrality, where all geodesic paths are considered equally important. The interaction networks depicted in Figure 4 clarify this point. In the interaction network of Figure 4a, where all genes between 2 and 7 are needed to connect the most associated genes 1 and 8, $\gamma$ centrality behaves similar to degree centrality, providing the same level of importance to all genes between 2 and 7 (differently from the other measures, which assign the biggest amount of importance to nodes 4 and 5 and the smallest amount to 2 and 7). On the other hand, we would tend to discard a long path between two genes, in favor of a one-edge path, because in this case it imposes additional intermediaries genes which are not needed to connect associated genes. This is the case of the interaction network depicted in Figure 4b, where genes 3 and 5 are intermediary genes not necessary to connect associated genes 1 and 8, and therefore they receive a null level of centrality both from $\gamma$ and betweenness centralities, whereas the other measures give an intermediate level of centralities to such nodes.

**Table 3.** Most 9 central genes according to $\gamma$ centrality

| Symbol | Name | Appr. $\gamma$ |
|---|---|---|
| UBE1$^{2,3}$ | ubiquitin-activating enzyme E1 | 0.0018 |
| BEXL1$^{3}$ | brain expressed X-linked-like 1 | 0.0014 |
| SFPQ | splicing factor proline/glutamine-rich | 0.0012 |
| SCD$^{3}$ | stearoyl-CoA desaturase (delta-9-desaturase) | 0.0010 |
| ODF4$^{1,2,3}$ | outer dense fiber of sperm tails 4 | 0.0010 |
| SPAG9$^{2,3}$ | sperm associated antigen 9 | 0.0010 |
| | (polypyrimidine tract binding protein associated) | 0.0010 |
| OR2B2 | olfactory receptor, family 2, subfam. B, memb. 2 | 0.0010 |
| STK23$^{1,2,3}$ | serine/threonine kinase 23 | 0.0010 |
| THRAP1 | thyroid hormone receptor associated protein 1 | 0.0008 |

Numbers shows genes found among the nine most central genes according to degree centrality ($^{1}$), closeness centrality ($^{2}$), betweenness centrality ($^{3}$) and eigenvector centrality ($^{4}$).



| gene | γ | Deg. | Clos. | Bet. | Eigen. |
|---|---|---|---|---|---|
| 1 | -0.38 | 1 | 0.25 | 0 | 0.35 |
| 2 | 0.13 | 2 | 0.32 | 6 | 0.65 |
| 3 | 0.13 | 2 | 0.39 | 10 | 0.88 |
| 4 | 0.13 | 2 | 0.44 | 12 | 1.00 |
| 5 | 0.13 | 2 | 0.44 | 12 | 1.00 |
| 6 | 0.13 | 2 | 0.39 | 10 | 0.88 |
| 7 | 0.13 | 2 | 0.32 | 6 | 0.65 |
| 8 | -0.38 | 1 | 0.25 | 0 | 0.35 |

| gene | γ | Deg. | Clos. | Bet. | Eigen. |
|---|---|---|---|---|---|
| 1 | -0.33 | 1 | 0.35 | 0 | 0.26 |
| 2 | 0.17 | 3 | 0.50 | 6 | 0.70 |
| 3 | 0.00 | 2 | 0.47 | 0 | 0.63 |
| 4 | 0.17 | 4 | 0.64 | 12 | 1.00 |
| 5 | 0.00 | 2 | 0.50 | 0 | 0.64 |
| 6 | 0.17 | 3 | 0.58 | 10 | 0.72 |
| 7 | 0.17 | 2 | 0.44 | 6 | 0.31 |
| 8 | -0.33 | 1 | 0.32 | 0 | 0.12 |

**Fig. 4.** Two different interaction networks (**A** and **B**) with eight genes (interactions are represented by thick lines). Gene 1 and 8 are the most associated genes in both networks, which directly interact (thin lines) to the key gene $k$. Centrality values of nodes according to different centrality measures are shown in the corresponding tables.

## 5 CONCLUSION

In this article, a new measure of the importance of genes in biological networks based on coalitional games is introduced. The new measure, calculated from the Shapley value of two coalitional games, has been used to express the centrality of each gene in interaction with the others and keeping into account a priori knowledge about genes playing a key function on a certain biological process.

The use of $\gamma$ index as a centrality measure is supported by the basic intuition that it is a difference of power indices between a

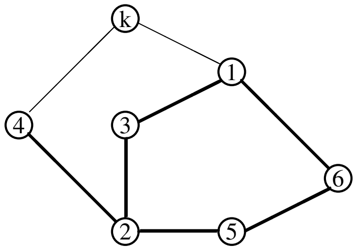


**Fig. 5.** An interaction network $\check{\Gamma}$ (thick lines) and the interactions of a gkg situation described with only one keygene $k$ (thin lines). Note that the subgraphs individuated by paths $(4,2,3,1)$ and $(4,2,6,5,1)$ satisfy the hypothesis (1), (2) and (3) of the TAM property. So, the TAM property would require that the sum of centrality values of nodes which are in $(4,2,3,1)$ is larger or equal than the sum of nodes in $(4,2,6,5,1)$. It is easy to check that, according to relations (4) and (8), $\gamma_3 = \frac{1}{4} - \frac{1}{6} = \frac{1}{12} > 2 \times (\frac{1}{5} - \frac{1}{6}) = \frac{1}{15} = \gamma_5 + \gamma_6$.

situation where binary interactions are considered (i.e. the graph-restriction game) and another one where they are not (i.e. the association game), and by the comparison with properties related to other centrality measures on some examples. In order to generalize these argumentations, an important issue for future research is to address a comprehensive analysis of the properties satisfied by the Shapley value on graph-restricted games, with the objective to better contextualize its interpretation as a centrality measure.

In this direction, we believe that the following property [namely, *Total Aggregation Monotonicity* (TAM) property] may be a crucial property both for the interpretation and the axiomatic characterization of the $\gamma$ centrality index. Consider a gkg situation $(N,\{k\},I)$ (i.e. with only one key gene) and an interaction network $\langle N,\Gamma\rangle$.

If there exist two subgraphs $\langle S,\Gamma^S\rangle$ and $\langle T,\Gamma^T\rangle$ such that:

(1) $S \cap T = W \supseteq N_k$ (the set of the most associated genes is a subset of the intersection of $S$ and $T$),

(2) $\Gamma^S \cup \Gamma^T = \Gamma$ (subgraphs are exhaustive: together they represent all the interactions in $\Gamma$),

(3) $|S| \leq |T|$ (the cardinality of $S$ is smaller than the cardinality of $T$),

then the TAM property requires that the sum of the centrality values given to nodes in $S$ is larger or equal than the sum of the centrality values given to nodes in $T$. The interpretation of the TAM property is related to the basic principle that the smaller pathways of genes provide a less complex explanation of the observed network of interactions, and for this reason they must be put into prominence. In addition, it may serve to regulate the behavior of a desired centrality index on separated components after the decomposition of an interaction network in simpler subgraphs. An example showing how the TAM property applies to $\gamma$ centrality is given in Figure 5.

An approximation method for the calculation of $\gamma$ centrality in practical biological networks is also presented. According to this procedure, the generation of all spanning forests in a biological network is not needed, but the analysis is limited to a smaller subset of forests characterized by the property that each pair of most associated genes is connected by a shortest path.

According to (Zhou *et al*., 2002), where a method of analysis using the shortest path to identify genes from the same biological process has been presented and validated on yeast interaction networks, a shortest path between two genes $i$ and $j$ involved in the same biological process represents the most reliable explanation of dependence between $i$ and $j$. As a consequence of their validation procedure, (Zhou *et al*., 2002) demonstrated that the genes on the shortest path between $i$ and $j$ are likely to be important intermediate players in the same process.

Therefore, it seems natural to focus on shortest paths in order to provide a more parsimonious representation of a graph-restricted game that makes feasible the application on a real interaction network. On the other hand, many shortest paths may exist between two genes on an interaction network, and therefore the selection of those genes that better represent the dependance between the most associated genes is still a problem involving the interaction of all possible coalitions, in the restricted domain defined by all shortest paths.

Of course, the price for using the method based on shortest paths is that genes outside those particular paths receive a null value of approximated $\gamma$ centrality, even if their exact $\gamma$ value is not null. On the other hand, this result is not in contrast with the principle that genes which are not on the shortest paths should receive a lower amount of centrality than genes on the shortest paths. In other words, the approximated method provides results which are consistent with the TAM property illustrated above.

## ACKNOWLEDGEMENTS

The authors are grateful to three anonymous referees for their comments, which helped to improve the quality of the article significantly.

*Conflict of Interest*: none declared.

## REFERENCES

Albino,D. *et al*. (2008) Identification of low intratumoral gene expression heterogeneity in neuroblastic tumors by wide-genome expression analysis and game theory. *Cancer*, **113**, 1412–1422.

Alon,N. (1990) The number of spanning trees in regular graphs. *Random Struct. Algorithms*, **1**, 175–181.

Amaratunga,D. and Cabrera,J. (2004) *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley-Interscience, New Jersey.

Bavelas,A. (1948) A mathematical model for small group structures. *Hum. Organ.*, **7**, 16–30.

Beauchamp,M. (1965) An improved index of centrality. *Behav. Sci.*, **10**, 161–163.

Bonacich,P. (1972) Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.*, **2**, 113–120.

Bondy,J.A. and Murty,U.S.R. (1976) *Graph Theory with Applications*. Macmillan, London.

Calderòn-Garcidueña,L. *et al*. (2007) Pediatric respiratory and systemic effects of chronic air pollution exposure: nose, lung, heart, and brain pathology. *Toxicol. Pathol.*, **35**, 154–162.

Carlson,M. *et al*. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, **7**, 40.

Clark,J. *et al*. (1997) Fusion of splicing factor genes PSF and NonO (p54nrb) to the TFE3 gene in papillary renal cell carcinoma. *Oncogene*, **15**, 2233–2239.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal*, Complex Systems, 1695. Available at http://igraph.sf.net (last accessed date August 19, 2010).

Elferink,J.G. and Ebbenhout,J.L. (1988) Asbestos-induced activation of the respiratory burst in rabbit neutrophils. *Res. Commun. Chem. Pathol. Pharmacol.*, **61**, 201–211.

Esteban,F.J. and Wall,D.P. (2009) Using game theory to detect genes involved in Autism Spectrum Disorder. *Top* [Epub ahead of print; doi: 10.1007/s11750-009-0111-6, July 24, 2009].

Fragnelli,V. and Moretti,S. (2008) A game theoretical approach to the classification problem in gene expression data analysis, *Comput. Math. Appl.*, **55**, 950–959.

Freeman,L.C. (1977) A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–41.

Gabow,H.N. and Myers,E.W. (1978) Finding all spanning trees of directed and undirected graphs. *SIAM J. Comput.*, **7**, 280–287.

Golub,T. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gòmez,D. *et al*. (2003) Centrality and power in social networks: a game theoretic approach. *Math. Soc. Sci.*, **46**, 27–54.

Gòmez,D.G. *et al*. (2004) Splitting graphs when calculating Myerson value for pure overhead games. *Math. Methods Oper. Res.*, **59**, 479–489.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Junker,B.H. *et al*. (2006) Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, **7**, 219.

Kapoor,S. and Ramesh,H. (1995) Algorithms for enumerating all spanning trees of undirected and weighted graphs. *SIAM J. Comput.*, **24**, 247–265.

Lucchetti,R. *et al*. (2009) The Shapley and Banzhaf indices in microarray games. *Comput. Oper. Res.*, **37**, 1406–1412.

Mason,O. and Verwoerd,M. (2007) Graph theory and networks in biology. *IET Syst. Biol.*, **12**, 89–119.

Minty,G.J. (1965) A simple algorithm for listing all the trees of a graph. *IEEE Trans. Circuit Theory*, **CT-12**, 120.

Moretti,S. (2009) Game theory applied to gene expression analysis. *4OR-Q. J. Oper. Research*, **7**, 195–198.

Moretti,S. (2010) Statistical analysis of the Shapley value for microarray games *Comput. Oper. Res.*, **37**, 1413–1418.

Moretti,S. *et al*. (2008) Combining Shapley value and statistics to the analysis of gene expression data in children exposed to air pollution. *BMC Bioinformatics*, **9**, 361.

Moretti,S. and Patrone,F. (2008) Transversality of the Shapley value. *Top*, **16**, 1–41.

Moretti,S. *et al*. (2007) The class of microarray games and the relevance index for genes. *Top*, **15**, 256–280.

Myerson,R.B. (1977) Graphs and cooperation in games. *Math. Oper. Res.*, **2**, 225–229.

Nieminen,J. (1974) On centrality in a graph. *Scand. J. Psychol.*, **15**, 322–336.

Owen,G. (1995) *Game Theory*, 3rd edn. Academic Press, San Diego.

Parmigiani,G. *et al*. (2003) The analysis of gene expression data: an overview of Methods and Software. In Parmigiani,G. *et al*. (eds) *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York, pp. 1–45.

R Development Core Team (2004) R: a language and environment for statistical. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org (last accessed date May 2010).

Rubin,J. and Sive,H. (2007) whitesnake/sfpq is required for cell survival and neuronal development in the Zebrafish Laura Anne Lowery. *Dev. Dyn.*, **236**, 1347–1357.

Sabidussi,G. (1966) The centrality index of a graph. *Psychometrika*, **31**, 581–603.

Shapley,L.S. (1953) A Value for n-Person Games. In Kuhn,H.W. and Tucker,A.W. (eds) *Contributions to the Theory of Games II. Annals of Mathematics Studies 28*. Princeton University Press, Princeton, pp. 307–317.

Shapley,L.S. and Shubik,M. (1954) A method for evaluating the distribution of power in a committee system. *Am. Polit. Sci. Rev.*, **48**, 787–792.

Shav-Tal,Y. *et al*. (2001) Nuclear relocalization of the premRNA splicing factor PSF during apoptosis involves hyperphosphorylation, masking of antigenic epitopes, and changes in protein interactions. *Mol. Biol. Cell*, **12**, 2328–2340.

Shaw,M.E. (1954) Group strucure and the behaviour of individuals in small groups. *J. Psychol.*, **38**, 139–149.

Storey,J.D. and Tibshirani,R. (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarray. In Parmigiani,G. *et al*. (eds) *The Analysis Of Gene Expression Data: Methods And Software*. Springer, NY, pp. 272–290.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

van Leeuwen,D. *et al*. (2008) Genomic analysis suggests higher susceptibility of children to air pollution. *Carcinogenesis*, **29**, 977–983.

Zhang,B. and Horvath,S. (2005) A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 17.

Zhou,X. *et al*. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.