Optimal map of the modular structure of complex networks

New Journal of Physics
The open-access journal for physics

Optimal map of the modular structure of complex networks

A Arenas[1,2,4], J Borge-Holthoefer[1], S Gómez[1] and G Zamora-López[3]

[1] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[2] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3] Interdisciplinary Center for Dynamics of Complex System, University of Potsdam, 14415 Potsdam, Germany
E-mail: alexandre.arenas@urv.cat, javier.borge@urv.cat, sergio.gomez@urv.cat and gorka_agnld@yahoo.es

New Journal of Physics **12** (2010) 053009 (18pp)
Received 4 January 2010
Published 6 May 2010
Online at http://www.njp.org/
doi:10.1088/1367-2630/12/5/053009

Abstract. The modular structure is pervasive in many complex networks of interactions observed in natural, social and technological sciences. Its study sheds light on the relation between the structure and the function of complex systems. Generally speaking, modules are islands of highly connected nodes separated by a relatively small number of links. Every module can have the contributions of links from any node in the network. The challenge is to disentangle these contributions to understand how the modular structure is built. The main problem is that the analysis of a certain partition into modules involves, in principle, as much data as the number of modules times the number of nodes. To confront this challenge, here we first define the contribution matrix, the mathematical object containing all the information about the partition of interest, and then we use truncated singular value decomposition to extract the best representation of this matrix in a plane. The analysis of this projection allows us to scrutinize the skeleton of the modular structure, revealing the structure of individual modules and their interrelations.

[4] Author to whom any correspondence should be addressed.

Contents

1. Introduction

The concept of the modular structure in real complex networks [1] is revolutionizing the understanding of the evolution of complex systems [2]. A lot of efforts have been devoted to its automatic detection [3]–[5]; however, very little is known as yet about the actual skeleton of the detected modules that build the network. This skeleton is likely to be relevant to understanding why physical processes in complex networks, such as synchronization [6], present emergent phenomena that are affected by the existence of topological barriers between modules. We still lack fundamental tools to anticipate these phenomena from a topological perspective. The current work is intended to provide network scientists with novel tools to screen the modular structure. The comprehension of modular structure in networks necessarily demands the analysis of the contribution of each one of its constituents (nodes) to the modules. Recently, Guimerà *et al* [7, 8] advanced on this issue, proposing two descriptors to characterize the modular structure: the z-score (a measure of the number of standard deviations a data point is from the mean of a data set) of the internal degree of each node in its module and the participation coefficient (P), defined as how the node is positioned in its own module and with respect to other modules. Given a certain partition, the plot of nodes in the z–P plane admits a heuristic tagging of the nodes' role. The success of this representation relies on a consistent interpretation of topological roles of nodes according to the specific data analyzed.

Here, we introduce a formalism to reveal the characteristics of networks at the topological mesoscale, where the representation of the network is viewed as a set of interconnected modules. We propose a method, based on linear projection theory, to study the modular structure in networks that enables a systematic analysis and elucidation of its skeleton. Firstly, we construct a matrix containing all the information about the modular structure, and secondly, we find an optimal dimensional reduction of the information contained in it. In particular, we present the optimal mapping of the information about the modular structure (in the sense of least squares)

in a two-dimensional (2D) space. The method has been applied to synthetic and real networks. The statistical analysis of the geometrical projections allows us to characterize the structure of individual modules and their interrelations in a unified framework.

The paper is structured as follows. In section 2, we present the motivation of the method and the main findings to interpret the outcome. In section 3, the method is illustrated using synthetic networks whose structure is controlled. Finally, in section 4, the method is tested in real networks and an explanation of the results is offered.

2. Projection of the modular structure

A complex network (weighted or unweighted, directed or undirected) can be represented by its graph matrix $\boldsymbol{W}$, whose elements W_{ij} are the weights of the connections from any node i to any node j. Assuming that a certain partition of the network into modules is available, we plan to analyze this coarse grained structure. Note that the partition can be obtained by various methods and that the method we propose based on modularity [3] is one possibility. The main object of our analysis is the contribution matrix $\boldsymbol{C}$, of N nodes to M modules. The rows of $\boldsymbol{C}$ correspond to nodes, and the columns to modules. The analysis of this matrix is the focus of our research. The elements $C_{i\alpha}$ are the number of links that node i dedicates to module α, and can be easily obtained as the matrix multiplication between W_{ij} and the partition matrix $\boldsymbol{S}$:

$$C_{i\alpha} = \sum_{j=1}^{N} W_{ij} S_{j\alpha}, \tag{1}$$

where if the node j belongs to module α, $S_{j\alpha} = 1$, otherwise $S_{j\alpha} = 0$. The goal is to reveal the structure of individual modules, and their interrelations, from the matrix $\boldsymbol{C}$. To this end, we propose to deal with the high dimensionality of the original data by constructing a 2D map of the contribution matrix, minimizing the loss of information in the dimensional reduction and making it more amenable to further investigation.

2.1. Singular value decomposition (SVD) of the modular structure

The approach developed here consists of the analysis of $\boldsymbol{C}$ using SVD [9]. It stands for the factorization of a rectangular N-by-M real (or complex) matrix as follows:

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\dagger}, \tag{2}$$

where $\boldsymbol{U}$ is a unitary N-by-N matrix, $\boldsymbol{\Sigma}$ is a diagonal N-by-M matrix and $\boldsymbol{V}^{\dagger}$ denotes the conjugate transpose of $\boldsymbol{V}$, an M-by-M unitary matrix. This decomposition corresponds to a rotation or reflection around the origin, a non-uniform scale represented by the singular values (diagonal elements of $\boldsymbol{\Sigma}$) and (possibly) change in the number of dimensions, and finally again a rotation or reflection around the origin. This approach and its variants have been extraordinarily successful in many applications [9], in particular for the analysis of relationships between a set of documents and the words they contain. In this case, the decomposition yields information about word–word, word–document and document–document semantic associations; the technique is known as latent semantic indexing [10] or latent semantic analysis [11]. Our scenario is quite similar to this, where nodes resemble words and modules resemble documents. We suggest that a similar approach will help to unravel the relations between nodes' contributions and modules of a certain partition.

2.2. An optimal 2D map of the modular structure of networks

A practical use of SVD is dimensional reduction approximation, also known as truncated singular value decomposition (TSVD). It consists in keeping only some of the largest singular values to produce a least squares optimal, lower rank order approximation (see the appendix). In the following, we will consider the best approximation of $\boldsymbol{C}$ by a matrix of rank $r = 2$.

The main idea is to compute the projection of the contribution of nodes to a certain partition (rows of $\boldsymbol{C}$, namely $\boldsymbol{n}_i$ for the ith node) into the space spanned by the first two left singular vectors, the projection space $\mathcal{U}_2$ (see the appendix). We denote the projected contribution of the ith node as $\tilde{\boldsymbol{n}}_i$. Given that the transformation is information preserving [12], the map obtained gives an accurate representation of the main characteristics of the original data, visualizable and, in principle, easier to scrutinize. Note that the approach we propose has essential differences with classical pattern recognition techniques based on TSVD such as principal components analysis (PCA) or, equivalently, Karhunen–Loeve expansions. Our data (columns of $\boldsymbol{C}$) cannot be independently shifted to mean zero without losing its original meaning; this restriction prevents the straightforward application of the mentioned techniques and also differentiates our work from the modern techniques for the analysis of gene expression patterns [13, 14].

The main problem in using SVD lies always in the interpretation of its outcome. The combination of data in the process makes difficult a direct comparison between input and output. To overcome this problem, we point out the following geometrical properties of the projection of the rows of $\boldsymbol{C}$ we have defined (see the appendix for a mathematical description):

(i) Every module α has an intrinsic direction $\tilde{\boldsymbol{e}}_\alpha$ in the projection space $\mathcal{U}_2$ corresponding to the line of the projection of its internal nodes (those that have links exclusively inside the module). We call these directions *intramodular projections*. This property is essential to discern among modules that are cohesive, in the sense that the majority of nodes project in this direction, from those modules that are not cohesive.

(ii) Every module α has a distinguished direction $\tilde{\boldsymbol{m}}_\alpha$ in the projection space $\mathcal{U}_2$ corresponding to the vector sum of the contributions of all its nodes. We call these directions *modular projections*. The modular projection is relevant when compared to the intramodular projection because their deviations inform about the tendency to connect with other modules. Note that e_α and m_α are equal only if the module is disconnected from the rest of the network.

(iii) Any node contribution projection $\tilde{\boldsymbol{n}}_i$ is a linear combination of intramodular projections, the coefficient of each one being proportional to the original contribution $C_{i\alpha}$ of links of the node i to each module α. This property comes from the linearity of the projection, and expresses the contribution of nodes to the modules to which they are connected.

Consequently, from (i) and (iii), we can classify nodes. Nodes with only internal links have a distance to the origin proportional to its degree (or strength). Nodes with internal and external links separate from the intramodular projection proportionally to their contributions to other modules. From (ii) we can classify modules. Modules that have close modular projections are more interrelated. These geometrical facts are the key to relate the outcome of TSVD and the original data in our problem, see figure 1.

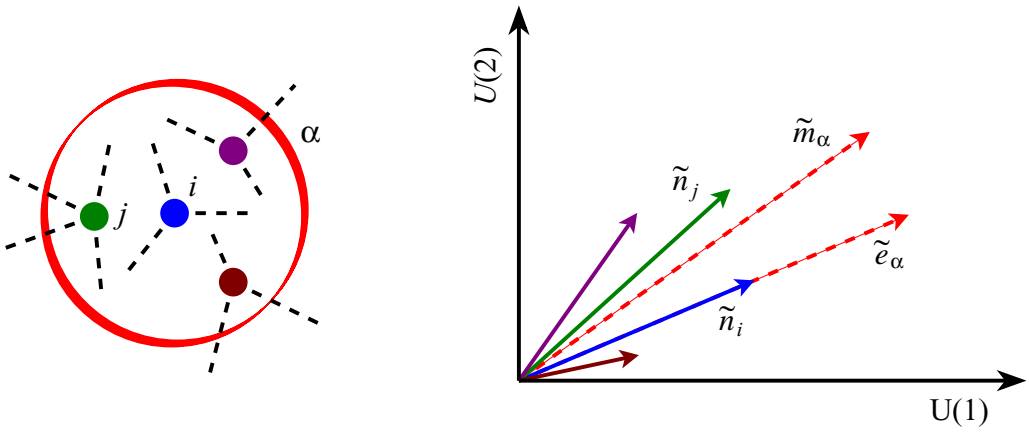

Figure 1. Geometrical scheme of the TSVD. The intramodular projection of module α, $\tilde{\boldsymbol{e}}_\alpha$ is the direction where all internal nodes lie (in the plot, node i). The node contribution projections $\tilde{\boldsymbol{n}}$ are represented by vectors in different colors. Finally, the modular projection $\tilde{\boldsymbol{m}}_\alpha$ is computed as the vector sum of all the node contribution projections belonging to it. Note that the intramodular projection and the modular projection do not coincide; the differences between both inform about the cohesiveness of the module.

2.3. Structure of individual modules

To study the structure of individual modules, we concentrate on the analysis of the projection of nodes' contributions in the plane $\mathcal{U}_2$. Keeping in mind the geometrical properties (i) and (iii) exposed above, we propose to extract structural information relative to each module by comparing the map of nodes' contributions to the intramodular projection directions. To this end, it is convenient to change to polar coordinates, where for each node i the radius R_i measures the length of its contribution projection vector $\tilde{\boldsymbol{n}}_i$, and θ_i the angle between $\tilde{\boldsymbol{n}}_i$ and the horizontal axis. We also define ϕ_i as the absolute distance in angle between $\tilde{\boldsymbol{n}}_i$ and the intramodular projection $\tilde{\boldsymbol{e}}_\alpha$ corresponding to its module α, i.e. $\phi_i = |\theta_i - \theta_{\tilde{\boldsymbol{e}}_\alpha}|$.

Using these coordinates R–ϕ, we find a way to interpret correctly the map of the contribution matrix in $\mathcal{U}_2$: (i) $R_{\text{int}} = R\cos\phi$ informs about the internal contribution of nodes to its corresponding module, as well as the contribution to its own module by connecting to others. To clarify the latter assertion, let us assume that a node i belonging to a module β has connections with the rest of the modules in the network. Given that this connectivity pattern is a linear combination of intramodular directions $\tilde{\boldsymbol{e}}_\alpha$, the vector sum implies that connecting with modules α having $|\theta_{\tilde{\boldsymbol{e}}_\beta} - \theta_{\tilde{\boldsymbol{e}}_\alpha}| > \pi/2$ decreases the module R and vice versa. (ii) $R_{\text{ext}} = R\sin\phi$ informs about the deviation (as the orthogonal distance) of each node to the contribution to its own module, see figure 2. It is also possible to study the spreading of ϕ by using other descriptors proposed in the context of synchronization [15].

We explore the internal structure of modules using the values of R_{int}, and we explore the boundary structure of modules using R_{ext}. Using descriptive statistics one can reveal and compare the structure of individual modules. Provided that the distribution of contributions is not necessarily Gaussian, an exploration in terms of z-scores is not convenient. Instead we use box-and-whisker charts for the variables, depicting the principal quartiles and the outliers (defined as having a value more than 1.5 IQR lower than the first quartile or 1.5 IQR higher than the third quartile, where IQR is the inter-quartile range).

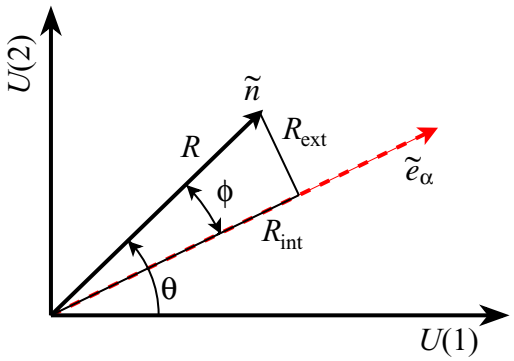

Figure 2. Schematic plot of the coordinates proposed to study the structure of individual modules. The relative distance of a node from its module is captured by the angle ϕ. The respective components R_{int} and R_{ext} are depicted.

The boxplots for the data of each module in the variable R_{int} allow for a visualization of the heterogeneity in the contribution of nodes building their corresponding modules, and an objective determination of distinguished nodes on its structure (outliers). Consequently, the boxplots in R_{ext} inform about the heterogeneity in the boundary connectivity. Nodes with links in only one module are not considered in these statistics because they do not provide relevant information about the boundaries (they have $\phi = 0$); only nodes that act as bridges between modules are taken into account. Considering internal nodes in these statistics would eventually produce a collapse of the quartiles to zero. Assuming that every module devotes some external links (otherwise they would be disconnected), the width of the boxes in this plot is proportional to the heterogeneity of such efforts. If only one node makes external connections, then the boxplot has zero width. Moreover, given two boxes equally wide, their position (median) determines which module contributes more to keeping the whole network connected.

2.4. Interrelations between modules

The analysis of the interrelations between modules is performed at the coarse grained level of its modular projections. The modular projections $\tilde{\boldsymbol{m}}_\alpha$ are aggregated measures of the nodes' contribution to their particular module. The normalized scalar product of modular projections provides a measure of the interrelations (overlapping) between different modules. A representation of these data in the form of a matrix ordered by the values of $\theta_{\tilde{\boldsymbol{m}}_\alpha}$ reveals the actual skeleton of the network at the topological mesoscale, see figure 3.

3. Application to synthetic networks

We start applying the methodology of analysis to synthetic networks, having control of the whole network structure. First, we analyze a network built from cliques of different sizes, and we consider a line of cliques from size 3 to 10, joined only by a unique link between them. We will consider two different partitions to test the method. The first partition consists of a module containing the larger clique and another containing the rest of the cliques, see figure 4(a). In the second partition, each clique forms a module, see figure 4(b). The plots in figures 4(c) and (d) (left) show the projections of the nodes' contributions in the plane spanned by the two first right

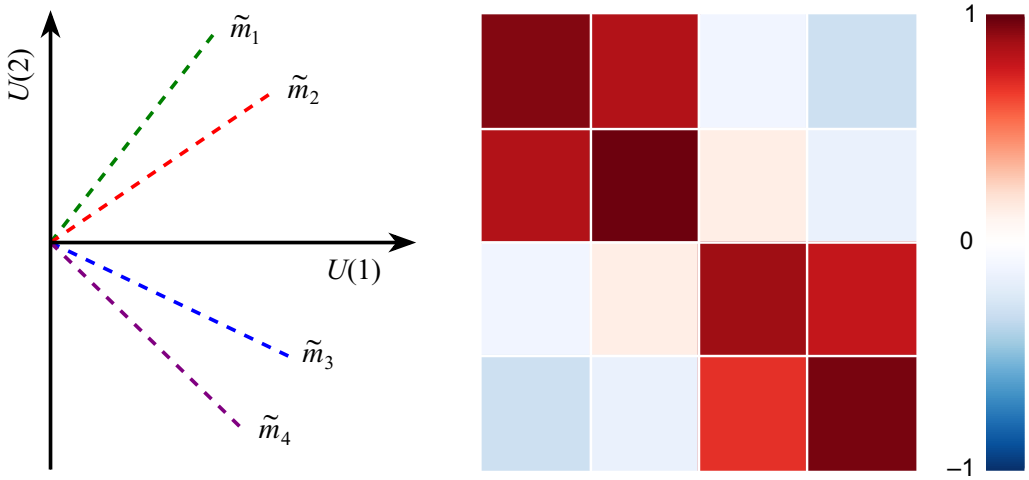

Figure 3. Schematic plot of the interrelation between the modular projections of four modules. The matrix represents the overlap computed as the scalar product between directions.

singular vectors $\mathcal{U}_2$, as well as the intramodular projections of each module in this plane. The data in $\mathcal{U}_2$ are transformed to polar coordinates for better visualization and simpler analysis, see figures 4(c) and (d) (right). The structure of these plots will be repeated in the next examples.

Projecting the contribution matrix corresponding to the partition in two modules figure 4(c), we observe clearly the relation between connectivity between nodes and the structure of both modules. The two distinguished nodes that connect both modules lie outside of the intramodular projections, while the rest of the nodes lie exactly in this direction. The different positions within the intramodular projections correspond to the degree of each node; nodes with identical contributions project to the same position. For the second partition, figure 4(d), the modules of size 3–9 are concentrated around a similar direction, while the clique of size 10 is separated from the rest. In the plot, we have zoomed in on the regions in the R–θ around the directions where nodes project. For every module the projection reflects two positions: one exactly in the intramodular direction corresponding to the internal nodes of the clique, and another corresponding to the node that acts as a connector with the following clique. The connectors towards the preceding clique (of lower size) are indistinguishable at the resolution of the plot, but also lie in a different direction.

Following the test, we now apply the method to a model of a network with a well-defined community structure that has been used as a benchmark for different community detection algorithms [5], proposed by Newman and Girvan [3]. In that model, the authors construct a network of 128 nodes as a set of four communities, each one formed by 32 nodes. Fixing the mean number of links per node at a value of 16, the parameter describing the sharpness of the community distribution is z_{in}, the average number of links within the community. A generalization of this model was proposed in [16] to include several hierarchical levels of communities. The hierarchy is defined as follows: we take a set of N nodes and divide it into n_1 groups of equal size; each of these groups is then divided into n_2 groups and so on, up to a number of steps k, which defines the number of hierarchical levels of the network. Then we add links to the networks in such a way that at each node we assign at random a number of z_1 neighbors within its group at the first level, z_2 neighbors within the group at the second level

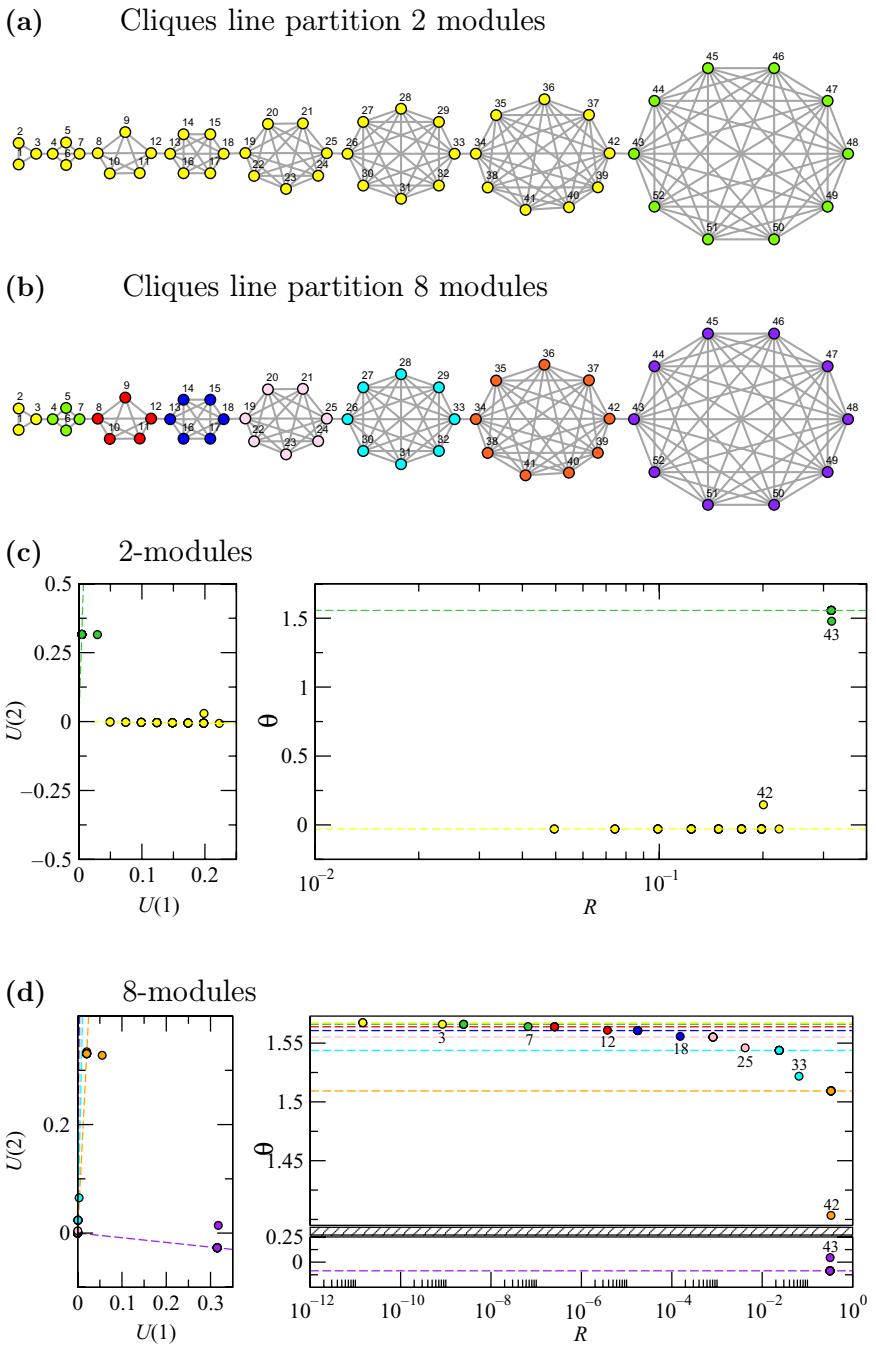

Figure 4. Optimal map of the modular structure for the optimal partition of the cliques network partitioned in two modules (a) and the cliques network partitioned in eight modules (b); each color corresponds to a different module of the given partition. In (c) and (d), we plot the projected space spanned by the two left singular vectors of the TSVD, $\mathcal{U}_2$ (left), and its transformation to polar coordinates R–θ (right), for each network. Dashed lines mark the directions of intramodular projections of each module. In (d) (right), we present a zoom in θ for better visual inspection.

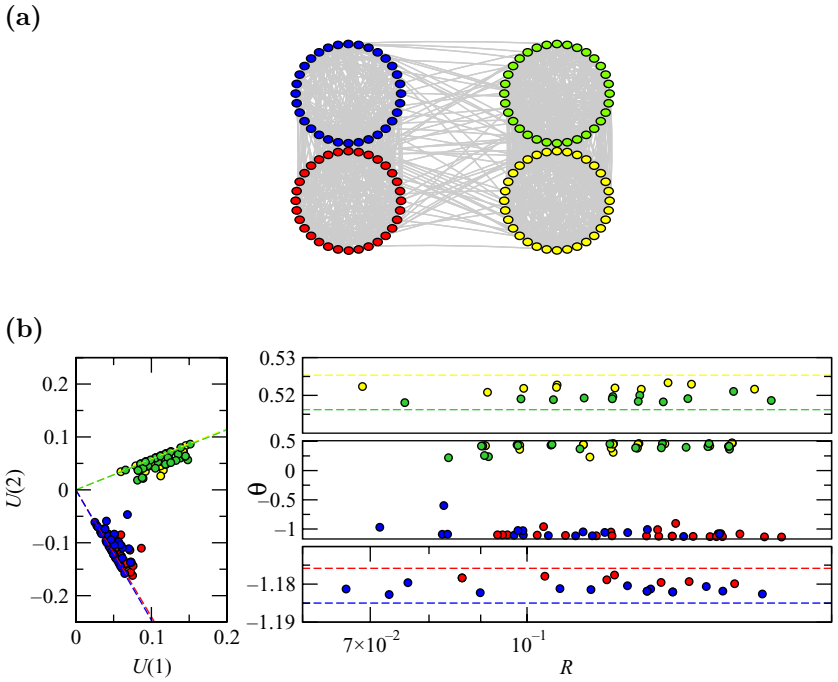

Figure 5. Analysis of a random homogeneous hierarchical network with community structure, see the text for details. (a) Network structure. (b) Projection as explained in figure 4.

and so on. There remains the number of links that each node has to the rest of the network; this we will call z_{out}. We construct a network with $N = 128$ nodes, two hierarchical levels with $n_1 = 2$, $n_2 = 2$, $z_1 = 5$, $z_2 = 10$ and $z_{\text{out}} = 1$. Again the method resolves the modular structure and individual contributions in the correct way, see figure 5. In appendix D, we also test the sensitivity and robustness of the method to slight changes in the predefined partition.

4. Application to real networks

The first network analyzed is the Zachary's karate club network [17], accounting for the study over two years of the friendship between 34 members of a karate club at a US university in 1970. The network in question was divided, at the end of the study period, into two groups after a dispute between the club's administrator (node 1) and the club's instructor (node 34), which ultimately resulted in the instructor leaving and starting a new club, taking about half of the original club's members with him. The partition we have used in our study corresponds to four modules resulting from optimizing modularity [3] using extremal optimization [18] and refined with Tabu search [19], providing a value of modularity $Q = 0.420$. After the projection, see figure 6, we observe nodes 1 and 3 in the green module and nodes 33 and 34 in the blue module, clearly distinguished by its value of R, denoting their important role in supporting the structure of both modules; however, they are not the nodes that connect with other modules. It is also remarkable that node 10 lies half-way between the modular directions of the larger modules assessing its unclassifiable nature (this node has been persistently misclassified by most of the community detection algorithms).

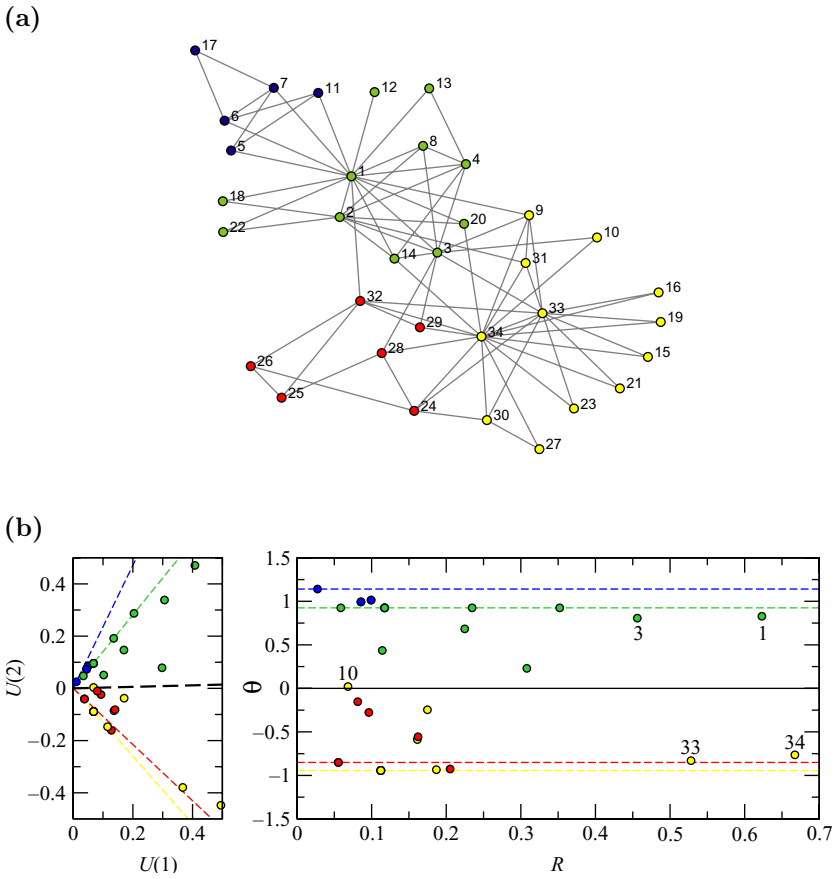

Figure 6. Analysis of the Zachary network for the four modules found by maximizing modularity. (a) The network with each module represented in a different color. (b) The projection as explained in figure 4.

The proposed mapping is also applied to two other real networks: the worldwide air transportation network and the AS–P2P Internet network. The airports network data set is composed of passenger flights operating in the time period 1 November 2000 to 31 October 2001, compiled by OAG Worldwide (Downers Grove, IL) and analyzed previously by Professor Amaral's group [8]. It consists of 3618 nodes (airports) and 14 142 links; we used the weighted network in our analysis. Airports corresponding to a metropolitan area have been collapsed into one node in the original database. The AS–P2P Internet data set considered is composed of autonomous systems (AS) [20] in the peer-to-peer (P2P) category, where two ASs freely exchange traffic between themselves and their customers, but do not exchange traffic from or to their providers or other peers [21]. We complemented this data set with the geographic localization of the ASs, resulting in 1217 nodes and 4058 links. We have optimized modularity [3] to find good partitions of the networks in modules. We have used the partition corresponding to 26 modules and modularity $Q = 0.649$ for the airports network, and 12 modules and $Q = 0.387$ for the AS–P2P network. Note that any partition, not necessarily the one corresponding to optimal modularity, can be analyzed as described.

The interesting aspect of applying the analysis to these two data sets is twofold: firstly, since both are geo-referenced, it is possible to assign a tag to each module corresponding to geographic areas, and secondly, the modular structure of both networks is substantially

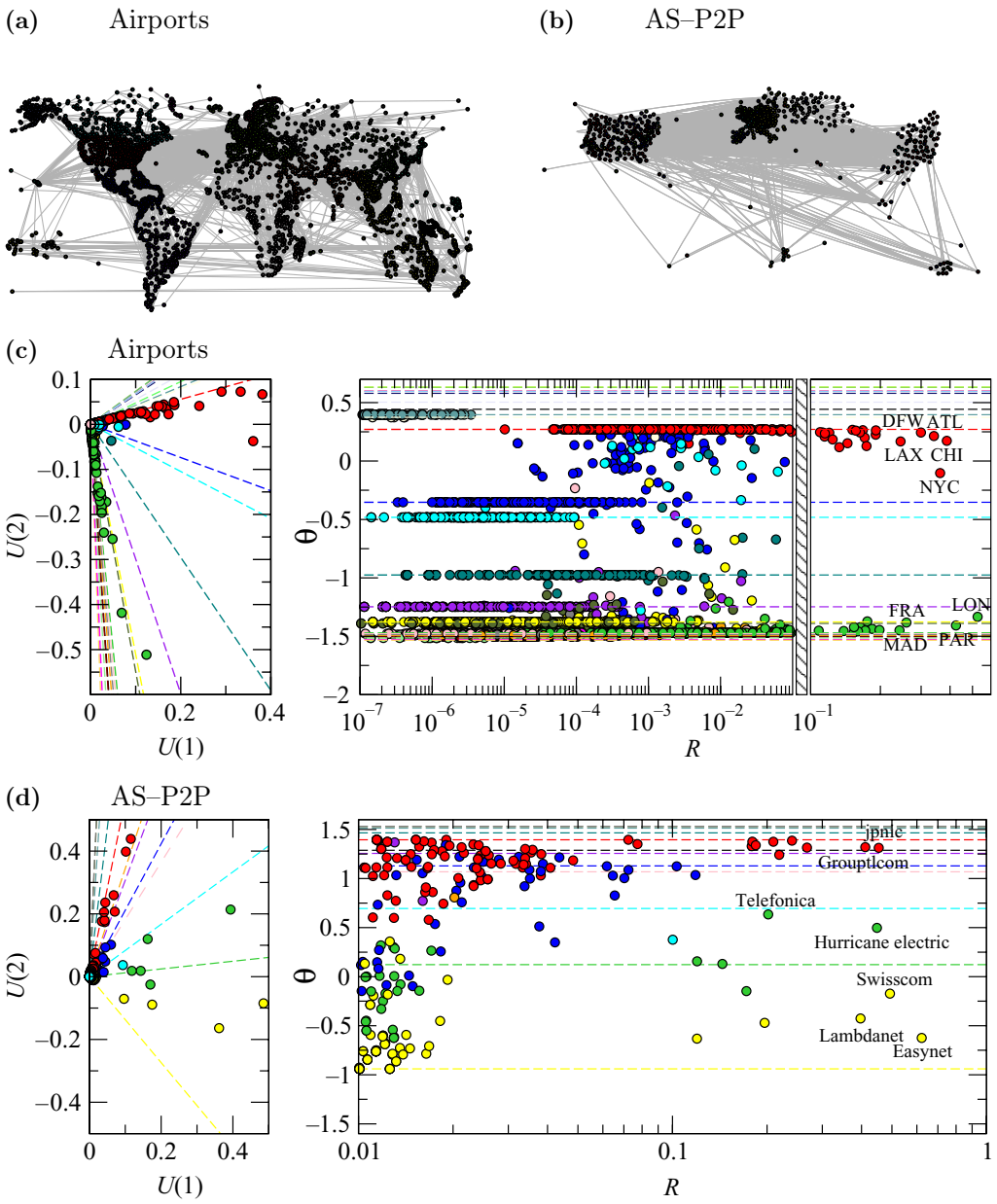

Figure 7. Optimal map of the modular structure for the optimal partition of the airports network (a) and the AS–P2P network (b); each color corresponds to a different module of the given partition. In (c) and (d), we plot the projected space spanned by the two left singular vectors of the TSVD, $\mathcal{U}_2$ (left), and its transformation to polar coordinates R–θ (right), for each network. Dashed lines mark the directions of intramodular projections of each module. Nodes whose contribution is totally internal to a module project exactly on their corresponding dashed lines. In the R–θ plot, we have labeled certain distinguished nodes that also correspond to very important airports and ASs in the world. For the airports network, we have magnified the area over 10^{-1} to identify the more important nodes in R. The loss of information associated with the 2D projection is 18.2% for the airports network and 15.8% for the AS–P2P network.

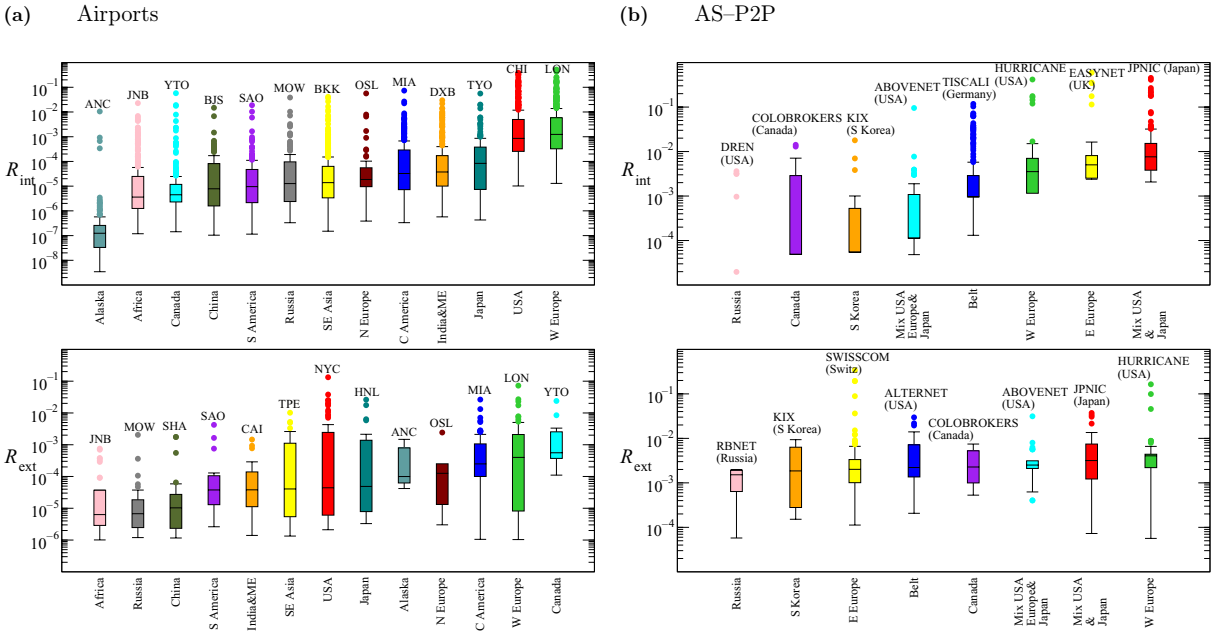

Figure 8. Box-and-whisker plots of R_{int} and R_{ext}, respectively, for the two networks depicted in figure 7. Modules are sorted according to medians in increasing order. We label the horizontal axis using names for the modules assigned according to the geographical location of at least the 75% of their nodes. We highlight whiskers and outliers in both networks. Only those modules whose structure is significant (more than 10 nodes) are represented in the plot.

different: while the airports network evolution has been mainly shaped by two well-defined continental blocks (USA and W Europe)[5], the AS–P2P network has been built in a more homogeneous way. It is very interesting to observe how the AS–P2P network, following a sort of 'wiring optimization', presents a community structure evenly distributed in areas covering a worldwide belt.

In figures 7(a) and (b), we plot the structure of the networks partitioned in modules; these conform to the original data that compose our contribution matrices. The geographical location has been added to the plot for visualization purposes but it has not been used in the analysis. The plots in figures 7(c) and (d) (left) show the projections of the nodes' contributions following the same structure of the precedent plots. The differences between both modular structures have clearly emerged in this projection: the airports network is basically polarized in two geographical areas, whereas in the AS–P2P network, this polarization does not exist. We also see how different airports and ASs excel in their values of R largely over the rest. This effect can be further developed by studying the structure of modules and their interrelations in each case.

The structure of modules is scrutinized in figure 8, where we depict the box-and-whisker plots of the internal contributions R_{int} and the external contributions R_{ext}. The results show the heterogeneity of each module of the partition. Remarkably, the method reveals outliers distinguished by their capability to support the internal structure of modules and also to

[5] We denote by N–S–E–W the four cardinal points North, South, East and West, respectively.

cross-connect them. In figure 8(a) (top), we observe that the USA and W Europe modules have medians greater than the percentiles-75 of the rest of the modules. This fact points out the extreme internal cohesion of both sites. We also observe that the lowest value in R_{int} median corresponds to Alaska; however, Anchorage leads the internal cohesion orders of magnitude beyond the core. In figure 8(a) (bottom), Canada, W Europe and C America provide the highest profile of boundary connectivity. Nevertheless, the role played by USA is still very significant because of its high percentiles and outliers. On the other side, Africa, Russia and China are less connected to the world than the rest of the modules. For the AS–P2P, the box-and-whisker plots of R_{int} in figure 8(b) (top) provide information about a slight dominance of the three modules E Europe, W Europe and the module containing USA and Japan. Here E Europe does not correspond to the political area but to a tag we use to represent a geographical area that is more oriental than the western, denoted as W Europe. In the case of R_{ext} in figure 8(b) (bottom), the similarity in range and medians reveals the homogeneity of the mesoscale of this network. Significantly, some highlighted ASs in the plot do not belong geographically to the assigned tag, although the major proportion of nodes in that module do (see E Europe, W Europe and Russia).

Finally, we plot the interrelations between modules in figure 9 by computing the scalar product of their respective modular projections. The labels of the matrix are chosen in decreasing order of the modular projection's angle $\theta_{\tilde{\boldsymbol{m}}_\alpha}$. For the airports network (figure 9(a)), we observe a clearly polarized structure in two main blocks, with a more diffuse central part overlapping both (corresponding to the communities mainly composed of nodes in Canada, Central America, Japan and South America). Japan is especially interesting for it maintains no preference in overlapping with any specific module in the network. In the AS–P2P network (figure 9(b)), we observe four groups, where neighbors in the analysis are in accordance with geographical neighbors. We remark that geographical information is not included in any part of the analysis; it simply emerges from the projection of the contribution matrix. The geographical correlation in the AS–P2P network could be surprising given that communities of use in P2P networks are related to contents or topics; however, many ASs have to pay other ASs to provide the connection between peers and then geopolitical constraints are revealed.

5. Conclusions

To summarize, firstly, we have reformulated the analysis of the modular structure, defining the object that contains all this information, and secondly, we apply SVD on this object. Dimensional reduction follows in a natural way from the properties of the truncation of SVD; in particular, we concentrate on the truncation of rank 2, with the idea of having a map of the modular structure amenable for analysis to any scientist. The approach is very simple and can be understood using basic algebra notions. The computational implementation is also affordable given the multiple software packages that include an automatic SVD (R and Matlab among others). The result is a formalism to study the skeleton of networks at the modular level. The most important problem we have faced in the current research was the interpretation of the outcome in terms of the original data. We have made a breakthrough in this interpretation by focusing our attention on the particular resulting geometry of the projected contribution of nodes. We also present a statistical analysis of the resulting map using the box-and-whisker plots based on percentiles, which are more appropriate than the use of z-scores that must assume a Gaussian distribution of values. Finally, we obtain the map of interrelations of the modular skeleton.

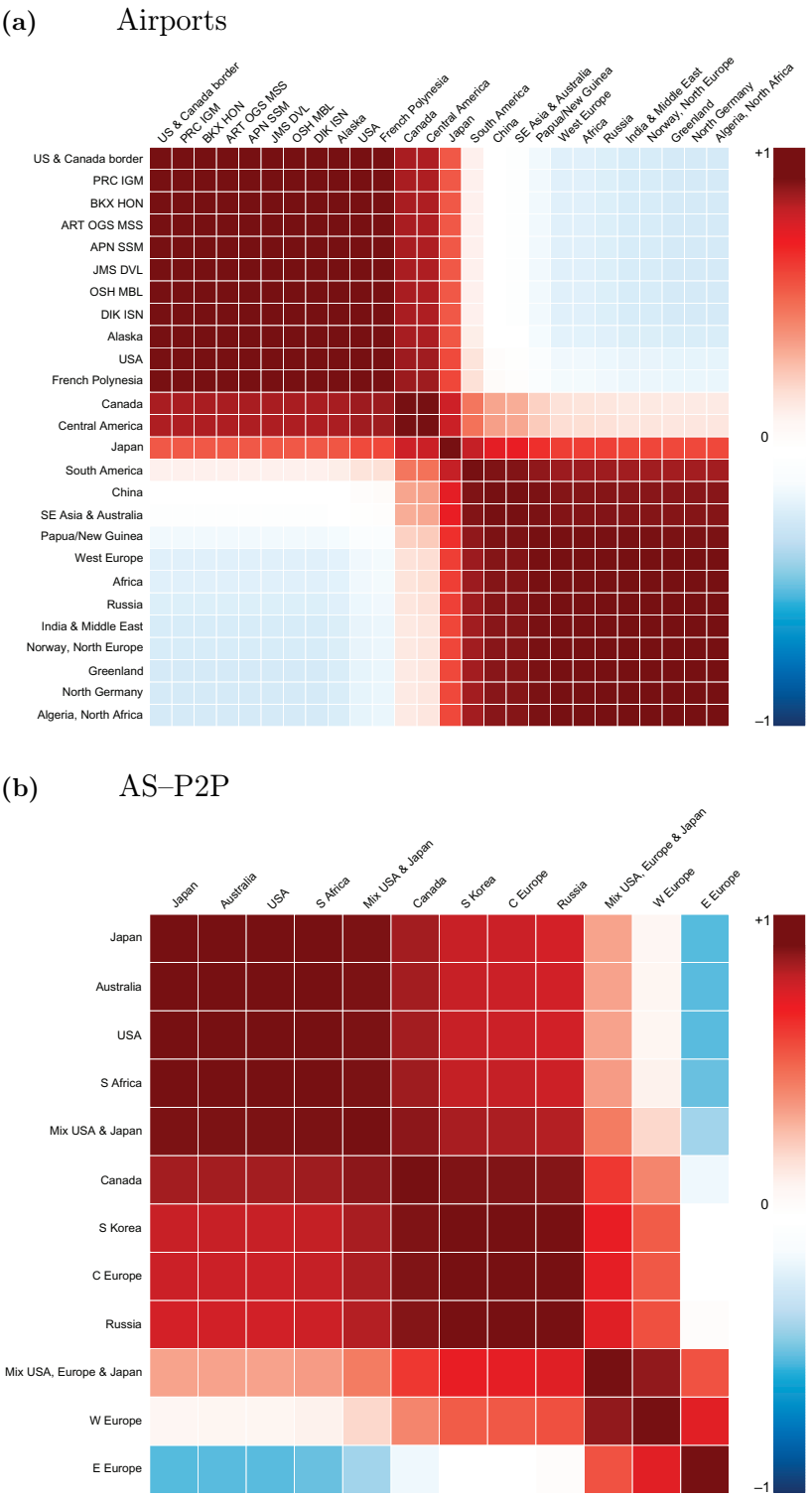

Figure 9. Overlap matrices between the modules composing the topological mesoscale of the networks plotted in figure 7. Each matrix corresponds to the normalized scalar product of the individual modular projections (see the text for details). Modules are sorted by decreasing order of the modular projection's angle in the plane $\mathcal{U}_2$.

The method proposed here might be very useful for scholars in different disciplines who seek access to an easy and tractable map of the complex empirical network data according to biological, functional or topological partitions. We suggest that the analysis of this map will be very helpful to anticipate the scope of dynamic emergent phenomena that depend on the structure and relations between modules. Spreading of viruses or synchronization processes are natural candidates to be analyzed by considering the organization of the map. Moreover, we suggest that the method can be used to graph bipartitioning by adaptively changing nodes between two modules while maximizing the angle in the R–θ plane between them. Further studies of the similarities between nodes' contribution projections can also help us to classify networks according to the role profiles of nodes [22] and/or modules.

Acknowledgments

We acknowledge A Díaz-Guilera, R Guimerà and C Zhou for useful discussions and the group of Professor L A N Amaral for sharing air transportation network data. This work was supported by the Spanish Ministry of Science and Technology (FIS2009-13730-C02-02) and the Generalitat de Catalunya (SGR-00838-2009). AA acknowledges the support of the Director, Office of Science, Computational and Technology Research, US Department of Energy, under contract number DE-AC02-05CH11231. GZ-L was supported by the Deutsche Forschungsgemeinschaft, research group FOR 868 (contract no KU 837/23-1) and the BioSim network of excellence, contract numbers LSHB-CT-2004-005137 and -65533.

Appendix A. Properties of TSVD

Let us assume that we preserve only the r largest singular values and neglect the remaining, substituting their value by zero; then the reduced matrix $\boldsymbol{C}_{\mathrm{r}} = \boldsymbol{U}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{V}^{\dagger}$ has several mathematical properties worth mentioning: firstly, it minimizes the Frobenius norm ($\|\boldsymbol{A}\|_F = \sqrt{\mathrm{trace}(\boldsymbol{A}\boldsymbol{A}^{\dagger})}$) of the difference $\|\boldsymbol{C} - \boldsymbol{C}_{\mathbf{r}}\|_F$, which means that among all possible matrices of rank r, $\boldsymbol{C}_{\mathbf{r}}$ is the best approximation in a least squares sense; secondly, $\boldsymbol{C}_{\mathbf{r}}$ is also the best approximation in the sense of statistics, as it maintains the most significant information portion of the original matrix [12]. The left and right singular vectors (from matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively) capture invariant distributions of values of the contribution of nodes to the different modules. In particular, the larger the singular value, the more the amount of information represented by their corresponding left and right singular vectors. We have used the LAPACK-based implementation of SVD in MATLAB. We warn that some numerical implementations of SVD suffer from a sign indeterminacy; in particular the one provided by MATLAB is such that the first singular vectors from an all-positive matrix always have all-negative elements, whose sign obviously should be switched to positive [23].

Appendix B. Projection using TSVD of rank 2

In the case of a rank $r = 2$ approximation, the unicity of the two-ranked decomposition is ensured [9] if the ordered singular values σ_i of the matrix $\boldsymbol{\Sigma}$ satisfy $\sigma_1 > \sigma_2 > \sigma_3$. This dimensional reduction is particularly interesting to depict results in a 2D plot for visualization purposes. In the new space, there are two different sets of singular vectors: the left singular

vectors (columns of matrix $\boldsymbol{U}$) and the right singular vectors (rows of matrix $\boldsymbol{V}^{\dagger}$). Given that we truncate at $r = 2$, we fix our analysis on the first two columns of $\boldsymbol{U}$; we call this the projection space $\mathcal{U}_2$. The coordinates $\tilde{\boldsymbol{n}}_i$ of the projection of the contributions $\boldsymbol{n}_i$ of node i are computed as follows:

$$\tilde{\boldsymbol{n}}_i = \boldsymbol{\Sigma}_2^{-1} \boldsymbol{V}^{\dagger} \boldsymbol{n}_i, \tag{B.1}$$

Here $\boldsymbol{\Sigma}_2^{-1}$ denotes the pseudo-inverse of the diagonal rectangular matrix $\boldsymbol{\Sigma}_2$ (singular values matrix truncated in two rows), simply obtained by inverting the values of the diagonal elements. It is possible to assess the loss of information of this projection compared to the initial data by computing the relative difference between the Frobenius norms:

$$E_r = \frac{||\boldsymbol{C}||_F - ||\boldsymbol{C}_{\mathbf{r}}||_F}{||\boldsymbol{C}||_F} = \frac{\sum_{\alpha=1}^{M} \sigma_\alpha^2 - \sum_{\alpha=1}^{r} \sigma_\alpha^2}{\sum_{\alpha=1}^{M} \sigma_\alpha^2}. \tag{B.2}$$

Appendix C. Geometrical properties of the projection of C

The intramodular projection $\tilde{\boldsymbol{e}}_\alpha$ corresponding to module α is defined as the projection of the Cartesian unit vector $\boldsymbol{e}_\alpha = (0, \ldots, 0, 1, 0, \ldots, 0)$ (the αth component is 1, the rest are zero), i.e.

$$\tilde{\boldsymbol{e}}_\alpha = \boldsymbol{\Sigma}_2^{-1} \boldsymbol{V}^{\dagger} \boldsymbol{e}_\alpha. \tag{C.1}$$

Any node in the original contribution matrix can be represented as

$$\boldsymbol{n}_i = \sum_{\alpha=1}^{M} C_{i\alpha} \boldsymbol{e}_\alpha. \tag{C.2}$$

Its projection gives the node contribution projection

$$\tilde{\boldsymbol{n}}_i = \sum_{\alpha=1}^{M} C_{i\alpha} (\boldsymbol{\Sigma}_2^{-1} \boldsymbol{V}^{\dagger} \boldsymbol{e}_\alpha) = \sum_{\alpha=1}^{M} C_{i\alpha} \tilde{\boldsymbol{e}}_\alpha, \tag{C.3}$$

a linear combination of intramodular projections. In particular, a node i whose contribution is totally internal to a module α is projected as $\tilde{\boldsymbol{n}}_i = k_i \tilde{\boldsymbol{e}}_\alpha$, where k_i is the node degree. The modular projections $\tilde{\boldsymbol{m}}_\alpha$ are computed as the vector sum of all the projections of node contributions, for those nodes belonging to module α, i.e.

$$\tilde{\boldsymbol{m}}_\alpha = \sum_{i=1}^{N} S_{i\alpha} \tilde{\boldsymbol{n}}_i. \tag{C.4}$$

Appendix D. Effect of noise on C

The method presented is pretty robust to perturbations in the partition or, equivalently, in the contribution matrix $\boldsymbol{C}$. To support the claim, we make the following experiment: using the benchmark network proposed by Girvan and Newman [1], see section 3, with 128 nodes, $z_{\text{in}} = 15$ and $z_{\text{out}} = 1$, we perform slight changes in the predefined partition, by moving nodes from module 1 to module 2. First we move only one node, then two nodes and finally eight

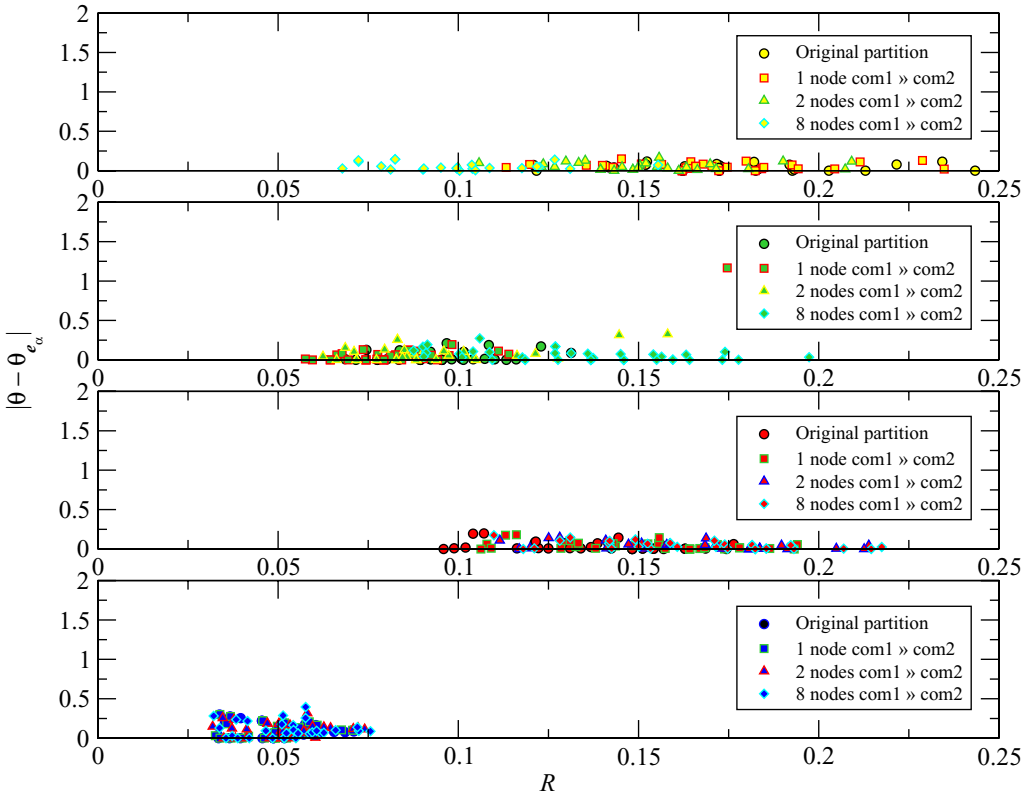

Figure D.1. Robustness of the method to noise in the partition. We show the separation from the intramodular directions of modules 1–4 (top to down) of all nodes; in particular, we track the deviation of the nodes when some of them have been assigned to the incorrect module. The nodes that have been moved are those that deviate more from the intramodular projection of module 2.

nodes. This changes matrix $\boldsymbol{C}$, which must in turn affect TSVD output. Figure D.1 contains the nodes' projection as the mentioned movements take place (squares, triangles and diamonds, respectively). Consistently, module 1's node projections progressively decrease in R. Module 2 balances this fact: it retains the weight leaving from module 1. Sensitivity to intermodular connections is also evidenced: when a single new node appears in module 2 (figure D.1, squares), ϕ_i has an outstanding value as compared to the rest; this is also evident when two nodes enter group 2 (figure D.1, triangles). When moving eight nodes, the effect is less drastic for the deviations in θ and more drastic in R. Unsurprisingly, modules 3 and 4 remain mostly unchanged; the interplay between modules 1 and 2 (nodes leaving from one group towards the other) does not drastically affect their internal characteristics, nor their importance in the whole structure.

References

[1] Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl Acad. Sci. USA* **99** 7821

[2] Vespignani A 2003 Evolution thinks modular *Nat. Genet.* **35** 118

[3] Newman M E J and Girvan 2004 Finding and evaluating community structure in networks *Phys. Rev.* E **69** 026113
[4] Palla G, Derényi I, Farkas I and Vicsek T 2005 Uncovering the overlapping community structure of complex networks in nature and society *Nature* **435** 814
[5] Danon L, Díaz-Guilera A, Duch J and Arenas A 2005 Comparing community structure identification *J. Stat. Mech.* P09008
[6] Arenas A, Díaz-Guilera A, Kurths J, Moreno Y and Zhou C 2008 Synchronization in complex networks *Phys. Rep.* **469** 93
[7] Guimerà R and Amaral L A N 2005 Functional cartography of metabolic networks *Nature* **433** 895
[8] Guimerà R, Mossa S, Turtschi A and Amaral L A N 2005 The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles *Proc. Natl Acad. Sci. USA* **102** 7794
[9] Golub G H and Van Loan C F 1996 *Matrix Computations* 3rd edn (Baltimore: Johns Hopkins University Press)
[10] Berry M W, Dumais S T and O'Brien G W 1995 Using linear algebra for intelligent information retrieval *SIAM Rev.* **37** 573
[11] Landauer T and Dumais S T 1997 A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge *Psychol. Rev.* **104** 211
[12] Chu M T and Golub G H 2005 *Inverse Eigenvalue Problems: Theory, Algorithms and Applications* (Oxford: Oxford University Press) pp 279–86
[13] Alter O, Brown P O and Botstein D 2000 Singular value decomposition for genome-wide expression data processing and modeling *Proc. Natl Acad. Sci. USA* **97** 10101
[14] Langfelder P and Horvath S 2007 Eigengene networks for studying the relationships between co-expression modules *BMC Syst. Biol.* **1** 54
[15] Rosenblum M G, Pikovsky A S, Kurths J, Schäfer C and Tass P 2001 Phase synchronization: from theory to data analysis *Handbook of Biological Physics, Neuro-informatics and Neural Modeling* ed F Moss and S Gielen (Amsterdam: Elsevier) p 279
[16] Arenas A, Díaz-Guilera A and Pérez-Vicente C 2006 Synchronization processes in complex networks *Physica* D **224** 27
[17] Zachary W W 1977 An information flow model for conflict and fission in small groups *J. Anthropol. Res.* **33** 452
[18] Duch J and Arenas A 2005 Community identification using extremal optimization *Phys. Rev.* E **72** 027104
[19] Arenas A, Fernández A and Gómez S 2008 Multiple resolution of the modular structure of complex networks *New J. Phys.* **10** 05039
[20] Dimitropoulos X, Krioukov D, Riley G Y and Claffy K C 2006 Revealing the autonomous system taxonomy: the machine learning approach *Passive and Active Measurements Workshop (PAM)* ed S Uhlig, K Papagiannaki and O Bonaventure (Berlin: Springer)
[21] Dimitropoulos X, Krioukov D, Fomenkov M, Huffaker B, Hyun Y, Claffy K C and Riley G 2007 AS relationships: inference and validation *Comput. ACM SIGCOMM. Commun. Rev.* **37** 29
[22] Guimerà R, Sales-Pardo M and Amaral L A N 2007 Classes of complex networks defined by role-to-role connectivity profiles *Nat. Phys.* **3** 63
[23] Bro R, Acar E and Kolda T G 2008 Resolving the sign ambiguity in the singular value decomposition *J. Chemometr.* **22** 135