PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Cite this article: Alvarez AJ, Sanz-Rodríguez CE, Cabrera JL. 2015 Weighting dissimilarities to detect communities in networks. *Phil. Trans. R. Soc. A* **373**: 20150108.
http://dx.doi.org/10.1098/rsta.2015.0108

Accepted: 19 August 2015

One contribution of 13 to a theme issue 'Topics on non-equilibrium statistical mechanics and nonlinear physics (II)'.

Subject Areas:
applied mathematics, complexity, computational biology, biochemistry

Keywords:
complex network, community detection, classification, social network analysis, metabolic network analysis, data mining

Author for correspondence:
Juan Luis Cabrera
e-mail: juluisca@gmail.com

Weighting dissimilarities to detect communities in networks

Alejandro J. Alvarez[1,2], Carlos E. Sanz-Rodríguez[1] and Juan Luis Cabrera[1]

[1]Stochastic Dynamics Laboratory, Center for Physics, Venezuelan Institute for Scientific Research, Caracas 1020-A, Venezuela
[2]Departamento de Física, FCFM, Universidad de Chile, Santiago, Chile

Many complex systems can be described as networks exhibiting inner organization as communities of nodes. The identification of communities is a key factor to understand community-based functionality. We propose a family of measures based on the weighted sum of two dissimilarity quantifiers that facilitates efficient classification of communities by tuning the quantifiers' relative weight to the network's particularities. Additionally, two new dissimilarities are introduced and incorporated in our analysis. The effectiveness of our approach is tested by examining the Zachary's Karate Club Network and the *Caenorhabditis elegans* reactions network. The analysis reveals the method's classification power as confirmed by the efficient detection of intrapathway metabolic functions in *C. elegans*.

1. Introduction

Characteristic inhomogeneities in real networks display order and organization, e.g. the local inhomogeneity in the distribution of links unveils the network organization in clusters of nodes. Such a feature is known as community structure [1]. In a community, nodes share some sort of similarity or common property. In a network, the communities may sustain functional meaning, e.g. in a social context clusters could be people with the same interests or buying patterns; in a biochemical network, clusters may perform specific functions such as energy production or storage. Despite the many works published on the topic, community detection in complex networks is still an active problem

THE ROYAL SOCIETY PUBLISHING

and has been addressed from different perspectives [2–18]. In this study, we follow a novel approach to this important problem. In particular, we propose a dissimilarity measure, $D_\epsilon(d_1, d_2)$, to detect communities. Such a measure is built with the weighted combination of two quantifiers, d_1 and d_2. This procedure is able to address different particularities of the network. To the best of our knowledge, there are no works dealing with weighted (dissimilarities) quantifiers in the context of community identification in complex networks.

2. The dissimilarity measure $D_\epsilon(d_1, d_2)$

Given the single quantifiers, d_1 and d_2, we define the parametric family of dissimilarity measures, $D_\epsilon(d_1, d_2)$, as the weighted sum

$$D_\epsilon(d_1, d_2) = \epsilon d_1 + (1 - \epsilon) d_2, \tag{2.1}$$

where $0 \leq \epsilon \leq 1$ is a weight parameter. In this expression, $D_\epsilon(d_1, d_2)$ is a matrix of components. We will focus on a set of specific dissimilarity quantifiers: link betweenness, the Jaccard dissimilarity index, Meet/Min and two new additional quantifiers that we introduce below. Let us consider the following definitions: given a network $\mathcal{N}$, nodes i and j, the betweenness centrality of an edge $\{i, j\}$ [19,20] is the sum of the fraction of all-pairs shortest paths that pass through $\{i, j\}$, i.e.

$$B_{ij} = \sum_{s,t \in \mathcal{N}} \frac{\sigma(s, t \mid \{i, j\})}{\sigma(s, t)}, \tag{2.2}$$

where $\sigma(s, t)$ is the number of shortest paths between nodes s and t, and $\sigma(s, t \mid \{i, j\})$ is the number of those paths passing through edge $\{i, j\}$.

If $N(i)$ denotes the neighbourhood nodes of node i, one of the simplest indexes, developed to compare regional floras [21], is the Jaccard dissimilarity index between nodes i and j, which is defined by

$$J_{ij} \equiv 1 - \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}. \tag{2.3}$$

The Meet/Min dissimilarity, mm_{ij}, as introduced by Golberg & Roth [22] and by Ravasz *et al.* [23] is expressed as

$$mm_{ij} \equiv 1 - \frac{|N(i) \cap N(j)|}{\min\{|N(i)|, |N(j)|\}}. \tag{2.4}$$

In this work, two additional quantifiers are introduced: (i) the Meet/Max, $MM(i, j)$, proposed as a variation of equation (2.5) and defined by

$$MM_{ij} \equiv 1 - \frac{|N(i) \cap N(j)|}{\mathrm{Max}\{|N(i)|, |N(j)|\}}, \tag{2.5}$$

and (ii) the intensity of interaction between two nodes, $\mathcal{I}_{ij}$, given by

$$\mathcal{I}_{ij} \equiv \begin{cases} 1 - \log_2\left(1 + \dfrac{1}{k_i k_j}\right) & \text{if } \{i, j\} \in \mathcal{E}(\mathcal{N}),\ k_i, k_j \neq 1 \\ 0 & \text{if } \{i, j\} \in \mathcal{E}(\mathcal{N}),\ k_i = 1 \text{ or } k_j = 1 \\ 1 & \text{in other cases,} \end{cases} \tag{2.6}$$

where k_i and k_j are the connectivity degrees of nodes i and j, respectively, and $\mathcal{E}(\mathcal{N})$ is the set of edges. Note the resemblance between $\mathcal{I}_{ij}$ and the channel capacity. $\mathcal{I}_{ij}$ quantifies information interchange between nodes, e.g. if only a link between two nodes occurs, the quantity $\mathcal{I}_{ij}$ is maximized because the flow of mutual information is not divided up by other nodes. On the contrary, if two high-degree nodes are connected the information circulating through them

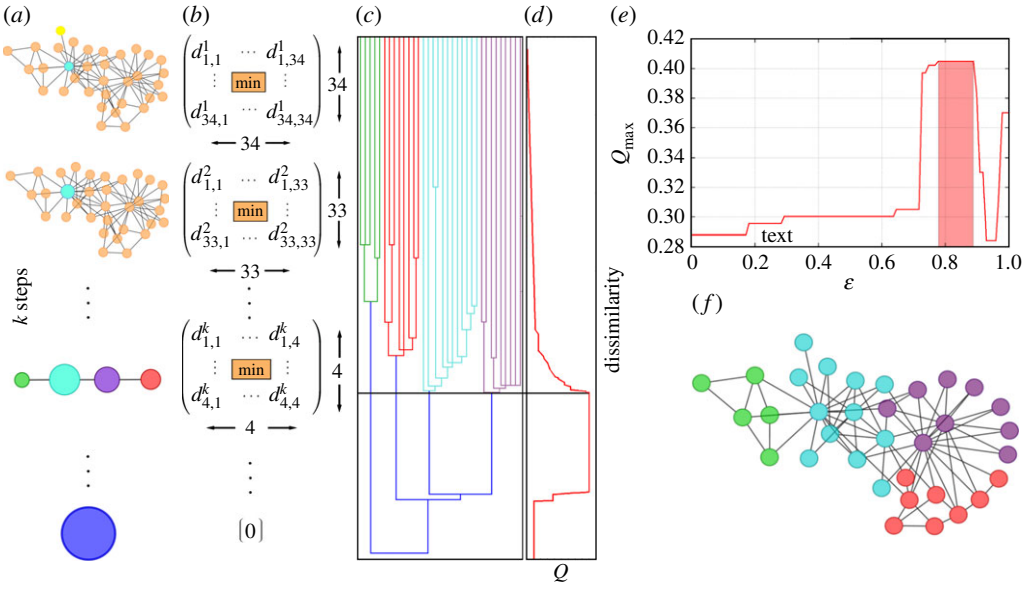


Figure 1. Schematic of the process of hierarchical clustering and ϵ tuning. Given a value of ϵ, at step $k = 0$ and (*a*) given a network (*b*) a matrix of dissimilarities $\hat{d}$ is calculated and its minimum entry determined. The corresponding indexes are collapsed into a single one as depicted by (*c*) the network's dendrogram showing the new network partition whose (*d*) modularity is now evaluated. This process is repeated recurrently through k steps until the network collapses into a single node. The maximum obtained value of the modularity, $Q_{\max}$, is calculated for each of the considered ϵ values as shown in (*e*). An optimal partition is then detected by the value (or set of values) of ϵ yielding the larger $Q_{\max}$ and (*f*) the best detected community structure. (Online version in colour.)

is shared with the additional neighbourhood nodes, producing poor effective communication between the two high-degree nodes. We found that this quantity is particularly useful to classify satellite nodes. As far as we know, $\mathcal{I}_{ij}$ has not been related to community detection yet. It seems to be one of the simplest dissimilarity measures to detect inner network communication.

Below, the following dissimilarity measures are evaluated:

$$D_\epsilon(J_{ij}, I_{ij}) = \epsilon J_{ij} + (1-\epsilon) I_{ij}, \tag{2.7}$$

$$D_\epsilon(mm_{ij}, I_{ij}) = \epsilon mm_{ij} + (1-\epsilon) I_{ij}, \tag{2.8}$$

$$D_\epsilon(MM_{ij}, I_{ij}) = \epsilon MM_{ij} + (1-\epsilon) I_{ij}, \tag{2.9}$$

$$D_\epsilon(\mathcal{B}_{ij}, I_{ij}) = \epsilon \mathcal{B}_{ij} + (1-\epsilon) I_{ij}, \tag{2.10}$$

$$D_\epsilon(J_{ij}, \mathcal{B}_{ij}) = \epsilon J_{ij} + (1-\epsilon) \mathcal{B}_{ij}, \tag{2.11}$$

$$D_\epsilon(mm_{ij}, \mathcal{B}_{ij}) = \epsilon mm_{ij} + (1-\epsilon) \mathcal{B}_{ij} \tag{2.12}$$

and

$$D_\epsilon(MM_{ij}, \mathcal{B}_{ij}) = \epsilon MM_{ij} + (1-\epsilon) \mathcal{B}_{ij}. \tag{2.13}$$

To carry out our approach, hierarchical clustering on the $D_\epsilon(d_1, d_2)$ matrix is developed, implementing a single-linkage renormalization or agglomerative method [24]. Consequently, the communities with the most similar nodes are recursively merged. This procedure is described in detail in figure 1. We have tested additional methods such as complete linkage [25] and average linkage [26], obtaining similar results. The performance of $D_\epsilon(d_1, d_2)$ was tested by analysing the community structure of: (i) the Zachary's Karate Club Network (ZKCN) [27] and (ii) the metabolic reactions network of *Caenorhabditis elegans*.

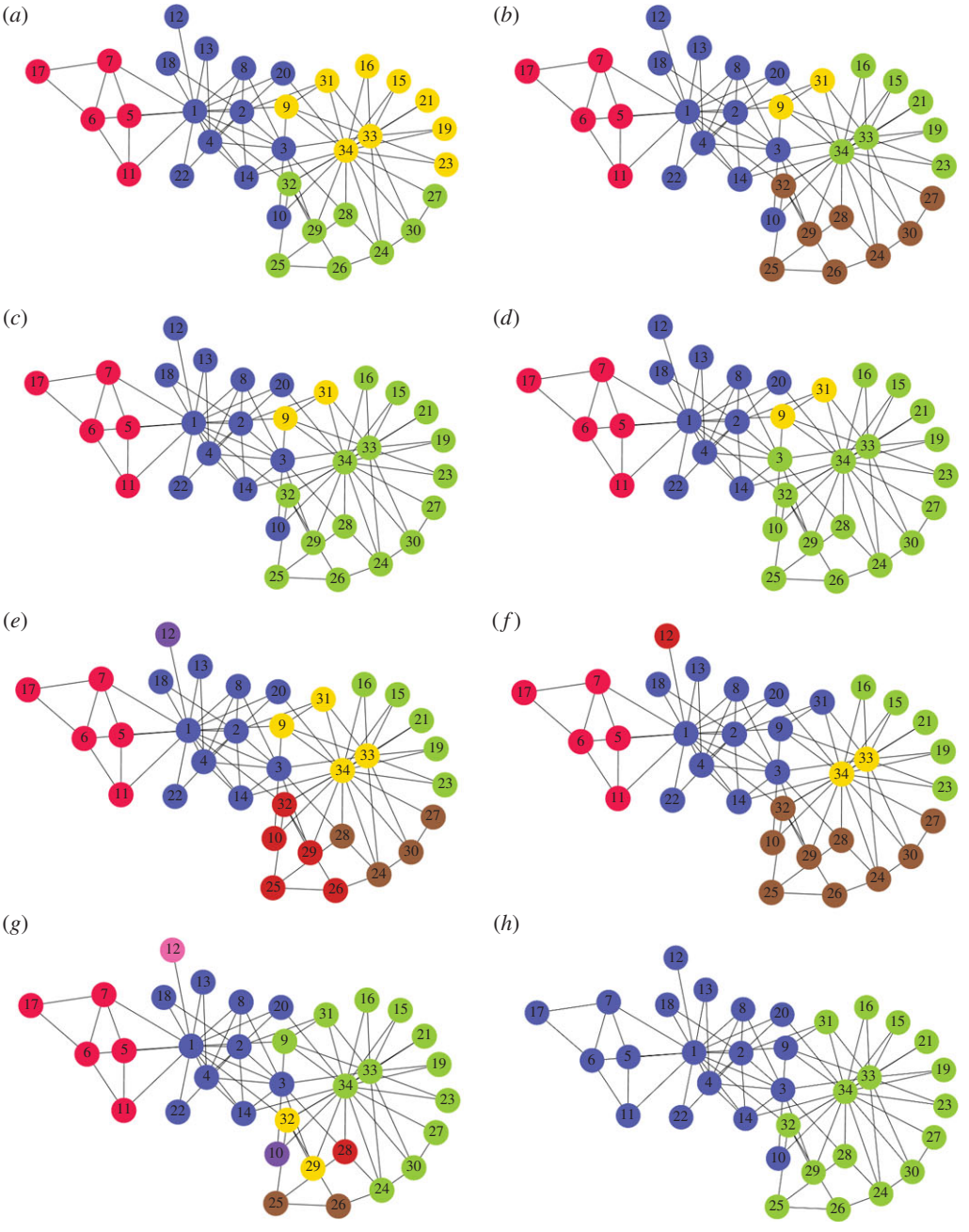


Figure 2. Community structures detected in the ZKCN using different quantifier combinations: (*a*) J_{ij} and $\mathcal{I}_{ij}$, (*b*) MM_{ij} and $\mathcal{I}_{ij}$, (*c*) mm_{ij} and $\mathcal{I}_{ij}$, (*d*) $\mathcal{B}_{i,j}$ and $\mathcal{I}_{ij}$, (*e*) J_{ij} and $\mathcal{B}_{i,j}$, (*f*) MM_{ij} and $\mathcal{B}_{i,j}$, (*g*) mm_{ij} and $\mathcal{B}_{i,j}$, (*h*) the two schools created after fission of the original club: the one led by the club's owner (blue) and the one led by the club's sensei (green). (Online version in colour.)

3. Zachary's Karate Club Network

ZKCN data were downloaded from the University of California Irvine Network Data Repository. Reported by Wayne Zachary while studying information flow patterns and fission of small social groups [27], the ZKCN is a network of friendship of 34 members of a karate club at a US university in the 1970s. ZKCN data were taken after the club split into two schools following an internal

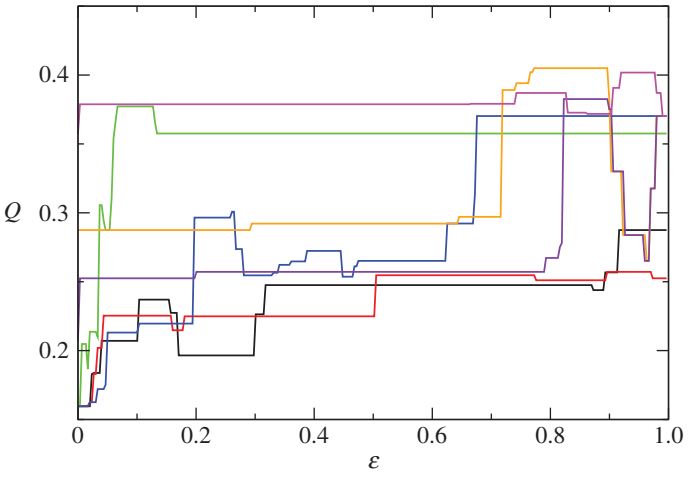


Figure 3. Modularity versus the parameter ϵ for each of the combinations of d_1 and d_2 according to equations (2.7)–(2.13) evaluated on the ZKCN. Different curves are for (orange) J_{ij} and $\mathcal{I}_{ij}$, equation (2.7) ($Q_{\max} = 0.405$); (magenta) mm_{ij} and $\mathcal{I}_{ij}$, equation (2.8) ($Q_{\max} = 0.402$); (violet) MM_{ij} and $\mathcal{I}_{ij}$, equation (2.9) ($Q_{\max} = 0.383$); (red) $\mathcal{B}_{i,j}$ and $\mathcal{I}_{ij}$, equation (2.10) ($Q_{\max} = 0.370$); (blue) J_{ij} and $\mathcal{B}_{i,j}$, equation (2.11) ($Q_{\max} = 0.287$); (green) mm_{ij} and $\mathcal{B}_{i,j}$, equation (2.12) ($Q_{\max} = 0.377$); and (black) MM_{ij} and $\mathcal{B}_{i,j}$, equation (2.13) ($Q_{\max} = 0.257$). The best results are obtained with the combination of the Jaccard and the intensity quantifiers on the interval $\epsilon \in [0.775, 0.897]$ (orange). (Online version in colour.)

dispute. Therefore, it is known that this network has at least two communities, i.e. the two schools created after fission: the one led by the club's owner and the one led by the club's sensei. The underlying community structure was determined by calculating $D_\epsilon(d_1, d_2)$ for different weighted combinations of the quantifiers d_1 and d_2. Results obtained with these combinations are displayed in figure 2*a–g*. A first obvious observation is that the identification of communities depends on the selection of the weight parameter ϵ and the particular measures d_1 and d_2. The classification in communities varies in composition. However, the identification of some particular communities, such as the one coloured bright pink and located on the left of the network's representation, does not depend on the implementation of $D_\epsilon(d_1, d_2)$. However, communities such as the blue one located at the network centre, while being quite stable in composition, vary slightly depending on the implemented quantifier. Meanwhile, there is a set of communities whose composition varies tremendously with the implementation of $D_\epsilon(d_1, d_2)$. The yellow, green, red and brown ones exemplify this situation. Finally, particular implementations of $D_\epsilon(d_1, d_2)$ determine isolated nodes. Certainly, the family of measures $D_\epsilon(d_1, d_2)$ show that, in all the analysed cases, the school led by node 1 (the club's owner) and the school led by node 34 (the club's sensei) may exhibit subdivisions into much more similar groups of nodes. The behaviour of the modularity values with the parameter ϵ for each of the evaluated combinations of d_1 and d_2 is shown in figure 3. In this representation, the orange line describes the best determined outcome, shown in figure 2*a*, that yields a modularity value $Q_{\max} = 0.405$. Such a measure was built with the weighted combination of the Jaccard and the interaction intensity given by equation (2.7) with optimum weight parameter values $\epsilon \in (0.778, 0.889)$. In such a case, four different communities are identified in the ZKCN.

4. *Caenorhabditis elegans* reactions network

Caenorhabditis elegans is the first multicellular organism whose genome was sequenced [28] and is widely used in genetics and neuroscience research. Metabolic network data from the work of Nerima *et al.* [29] were used here. The network of reactions was obtained by the projection of the metabolic network using the methods outlined by Newman *et al.* [30]. On this reactions network, $D_\epsilon(d_1, d_2)$ was calculated using the combinations given by equations (2.7)–(2.13). The

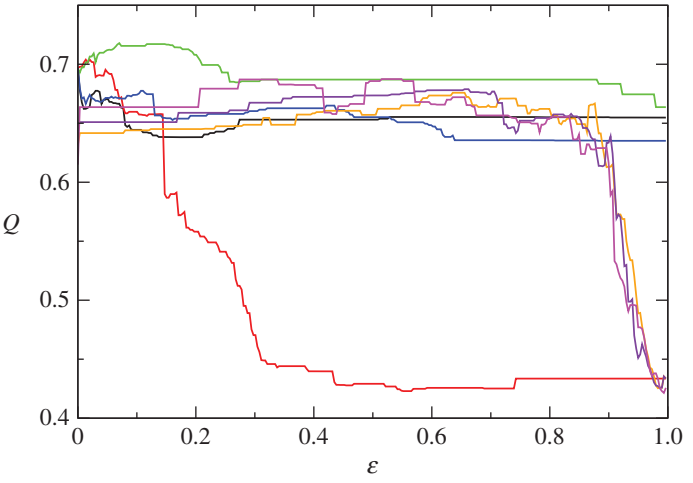
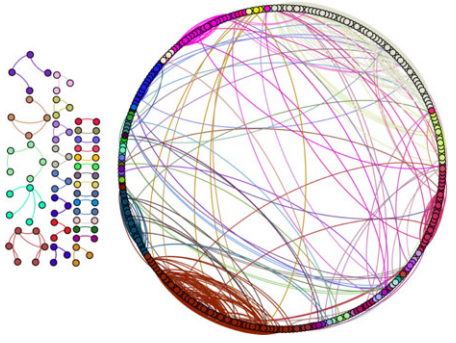


Figure 4. Modularity versus the parameter ϵ for each of the combinations of d_1 and d_2 according to equations (2.7)–(2.13) evaluated on the *C. elegans* reactions network. Different curves are for (orange) J_{ij} and $\mathcal{I}_{ij}$, equation (2.7) ($Q_{\max} = 0.676$); (magenta) mm_{ij} and $\mathcal{I}_{ij}$, equation (2.8) ($Q_{\max} = 0.687$); (violet) MM_{ij} and $\mathcal{I}_{ij}$, equation (2.9) ($Q_{\max} = 0.679$); (red) $\mathcal{B}_{i,j}$ and $\mathcal{I}_{ij}$, equation (2.10) ($Q_{\max} = 0.704$); (blue) J_{ij} and $\mathcal{B}_{i,j}$, equation (2.11) ($Q_{\max} = 0.697$); (green) mm_{ij} and $\mathcal{B}_{i,j}$, equation (2.12) ($Q_{\max} = 0.717$); and (black) MM_{ij} and $\mathcal{B}_{i,j}$, equation (2.13) ($Q_{\max} = 0.697$). The best results are obtained with the combination of the Meet/Min and betweenness quantifiers for $\epsilon = 0.071$ and the interval $\epsilon \in [0.117, 0.147]$ (green). (Online version in colour.)

(*a*)



(*b*)

community label	metabolic function
	amino acid degradation
	4-aminobutyrate metabolism
	biosynthesis of amino acids
	arginine and proline metabolism
	synthesis of deoxynucleotide
	glutathione metabolism
	glycine, serine and threonine metabolism
	biosynthesis of lysine amino acyl tRNA
	degradation of lysine
	biosynthesis of methionyl-tRNA
	non-oxidative pentose phosphate pathway and enzyme-associated
	biosynthesis of phenylalanine
	purine and pyrimidine metabolism interception with amino acids
	purine and pyrimidine metabolism
	purine degradation
	biosynthesis of putrescine from arginine
	biosynthesis of pyrimidine
	biosynthesis of spermidine
	tryptophan metabolism
	tyrosine and phenylalanine metabolism
	tyrosine and dopamine metabolism

Figure 5. (*a*) Best obtained community structure for the reactions network of *C. elegans* determined with the measure $D_\epsilon(d_1, d_2)$ built with a weighted combination of the Meet/Min and betweenness quantifiers given by equation (2.12). (*b*) Colour code for the metabolic functions associated with each detected community. (Online version in colour.)

best community structure was obtained with the weighted combination of the Meet/Min and betweenness centrality quantifiers given by equation (2.12) for $\epsilon = 0.071$ and the interval $\epsilon \in [0.117, 0.147]$ (green line of figure 4 yielding a value of $Q_{\max} = 0.717$ and $N_c = 74$ communities). The network community structure obtained with $Q_{\max}$ can be seen in figure 5. In figure 6, the resulting network's structure, the values of $Q_{\max}$ and the number of communities obtained with the rest of evaluations of $D_\epsilon(d_1, d_2)$ is reported.

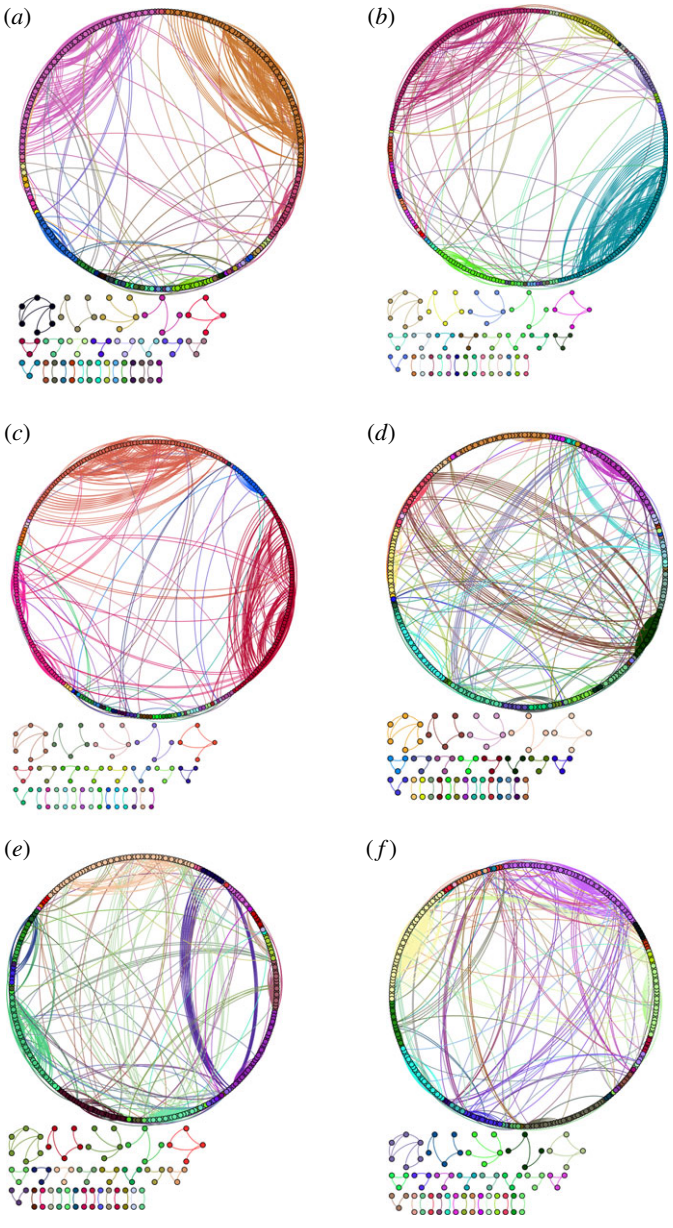


Figure 6. (*a*) Community structure of the reactions network of *C. elegans* determined with the measure $D_\epsilon(d_1, d_2)$ built with (*a*) J_{ij} and $\mathcal{I}_{ij}$, equation (2.7) ($Q_{\max} = 0.676$ and $N_c = 78$); (*b*) mm_{ij} and $\mathcal{I}_{ij}$, equation (2.8) ($Q_{\max} = 0.687$ and $N_c = 65$), (*c*) MM_{ij} and $\mathcal{I}_{ij}$, equation (2.9) ($Q_{\max} = 0.679$ and $N_c = 72$), (*d*) $\mathcal{B}_{i,j}$ and $\mathcal{I}_{ij}$, equation (2.10) ($Q_{\max} = 0.704$ and $N_c = 67$), (*e*) J_{ij} and $\mathcal{B}_{i,j}$, equation (2.11) ($Q_{\max} = 0.697$ and $N_c = 72$); and (*f*) MM_{ij} and $\mathcal{B}_{i,j}$, equation (2.13) ($Q_{\max} = 0.697$ and $N_c = 72$). Note that the colour codes are not the same as in figure 5. (Online version in colour.)

The functional significance of the detected structures using equation (2.12) was checked by searching in the KEGG database and manually curated for functions associated with the reactions at a given community. As expected our method classifies a set of reactions as communities developing specific metabolic pathways according to their gene ontology (GO). Remarkably, $D_\epsilon(d_1, d_2)$ is able to detect subtle differences in the same GO group. This is the case for the pentose–phosphate pathway, which is commonly grouped as a single pathway but in our case

we can separate it into its oxidative and non-oxidative component pathways. This is also true for the pathways of purine degradation and purine and pyrimidine metabolism.

5. Conclusion

Dissimilarity-based hierarchical clustering methods were conceived to account only for the quantitative or qualitative characteristics for which they were designed, being unable to access other aspects. Measures of dissimilarity were originally inspired by clustering of spatial (vectorial) data with no connections, in a context where communication or information flow were of no interest. As an example, consider the case of comparing two nodes' shared neighbourhoods. Obviously, such a procedure would not consider details such as information flow or connectivity. While single dissimilarity approaches have already been used to detect communities [31], it may seem insufficient to capture the intrinsic features of a complex network. A single quantifier does not capture the structural complexity. In such a context, in this work we propose a novel strategy that complements a quantifier's weaknesses with the strengths of another. As figure 2 illustrates, a particular measure $D_\epsilon(d_1, d_2)$ may perform better, depending on how well their constituent quantifiers, d_1 and d_2, adapt to the network's specificity. Thus, choosing an optimal measure may depend on the topological and statistical heterogeneity of the network under analysis. Adaptively coupling two different quantifiers may improve community detection while considering more than one relevant network's feature, e.g. local or global properties. The resulting network maximal modularity can be evaluated as a function of the parameter ϵ, thus tuning the weights in $D_\epsilon(d_1, d_2)$ to the best structural result. Such a tuning is not an artefact, as the resulting classification shows. In fact, it is possible to observe a good classification of the different communities of ZKCN. Remarkably, for the case of *C. elegans*, the method precisely separates all the metabolic pathways and intrapathway functions. Consequently, such a detailed function identification represents an improvement with respect to the GO classification available at KEGG.

In our paper, we explored two archetypal systems used widely in the literature to test methods for community detection in complex networks. In particular, we found that our method is useful to detect communities in sparse networks such as the metabolic network discussed above. In fact we found that weighting dissimilarities perform better than several of the methods recently published [8–18]. An analysis of larger networks would offer further insights into the dissimilarity's performance; however, such an analysis is outside the scope of this work and is left for a future publication. However, we analysed a number of different networks. We observed that when a network with more diverse characteristics is analysed the method adapts better and the modularity improves, as observed in our analysis of ZKCN ($Q_{\max} = 0.405$) and *C. elegans* ($Q_{\max} = 0.717$). In particular, our analysis considered different networks whose results are not included here but will be reported elsewhere. Such analysis includes (but is not limited to) the networks of *Candida albicans* ($Q_{\max} = 0.733$), *Dictyostelium discoideum* ($Q_{\max} = 0.744$), *Cryptosporidium hominis* ($Q_{\max} = 0.549$), *Schizosaccharomyces pombe* ($Q_{\max} = 0.751$) and *Entamoeba histolytica* ($Q_{\max} = 0.776$).

The current approach could be generalized to more than two measures, i.e. taking into account n different features to build a new measure

$$D_{\epsilon_k}(d_1, \ldots, d_n) = \sum_{k=1}^{n} \epsilon_k d_k, \tag{5.1}$$

where $d_1, \ldots, d_n$ are dissimilarity quantifiers capturing most of the features of a community in a complex network and $\sum_{k=1}^{n} \epsilon_k \equiv 1$. At this point, it must be remarked that the current method is not based on the optimization of modularity but that such a function is used as a quantifier providing selection criteria for the best cutting level on the dendrogram. As a result, the method does not suffer from a resolution limit. Moreover, variations of our approach can be implemented using different modularity functions, e.g. using the recently reported surprise [32], and possibly changing the criteria on the dendrogram's cut-off level. It seems obvious that a non-supervised version can be formulated with little effort.

Insights into the intermediation role that some nodes may play can be gained by integrating the information obtained with different implementations of $D_\epsilon(d_1, d_2)$. The application of such a procedure could reveal vertices with robust memberships and vertices playing a clear intermediate role. This aspect is left for future work.

Authors' contributions. A.J.A. and J.L.C. conceived and designed the research. A.J.A. developed the software. A.J.A. and C.E.S.-R. performed the numerical simulations. All authors participated in data analysis. A.J.A. and C.E.S.-R. prepared figures. J.L.C. coordinated the study. J.L.C. and A.J.A. wrote the manuscript. All authors gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. A.J.A. acknowledges an IVIC graduate fellowship. This work was supported by grant IVIC-141.

References

1. Girvan M, Newman MEJ. 2002 Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826. (doi:10.1073/pnas.122653799)
2. Newman MEJ. 2004 Detecting community structure in networks. *Eur. Phys. J. B* **38**, 321–330. (doi:10.1140/epjb/e2004-00124-y)
3. Danon L, Duch J, Arenas A, Díaz-Guilera A. 2007 Community structure identification. In *Large scale structure and dynamics of complex networks: from information technology to finance and natural science* (eds G Caldarelli, A Vespignani), pp. 93–114. Hackensack, NJ: World Scientific Publishing Co.
4. Schaeffer SE. 2007 Graph clustering. *Comp. Sci. Rev.* **1**, 27–64.
5. Fortunato S, Castellano C. 2009 Community structure in graphs. In *Encyclopedia of complexity and systems science* (ed. RA Meyers), pp. 1141–1163. New York, NY: Springer.
6. Porter MA, Onnela J-P, Mucha PJ. 2009 Communities in networks. *Not. Am. Math. Soc.* **56**, 1082–1097, 1164–1166.
7. Fortunato S. 2010 Community detection in graphs. *Phys. Rep.* **486**, 75–174. (doi:10.1016/j.physrep.2009.11.002)
8. Ahn Y-Y, Bagrow JP, Lehmann S. 2010 Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764. (doi:10.1038/nature09182)
9. Decelle A, Krzakala F, Moore C, Zdeborová L. 2011 Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Lett.* **107**, 065701. (doi:10.1103/PhysRevLett.107.065701)
10. Nadakuditi RR, Newman MEJ. 2012 Graph spectra and the detectability of community structure in networks. *Phys. Rev. Lett.* **108**, 188701. (doi:10.1103/PhysRevLett.108.188701)
11. Zhang Z-Y, Sun K-D, Wang S-Q. 2013 Enhanced community structure detection in complex networks with partial background information. *Sci. Rep.* **3**, 3241. (doi:10.1038/srep03241)
12. Malliaros FD, Vazirgiannis M. 2013 Clustering and community detection in directed networks: a survey. *Phys. Rep.* **533**, 95–142. (doi:10.1016/j.physrep.2013.08.002)
13. Sobolevsky S, Campari R, Belyi A, Ratti C. 2014 General optimization technique for high-quality community detection in complex networks. *Phys. Rev. E* **90**, 012811. (doi:10.1103/PhysRevE.90.012811)
14. De Meo P, Ferrara E, Fiumara G, Provetti A. 2014 Mixing local and global information for community detection in large networks. *J. Comp. Syst. Sci.* **80**, 72–87. (doi:10.1016/j.jcss.2013.03.012)
15. Subelj L, Bajec M. 2014 Group detection in complex networks: an algorithm and comparison of the state of the art. *Physica A* **397**, 144–156. (doi:10.1016/j.physa.2013.12.003)
16. Gong M, Liu J, Ma L, Qing C, Jiao L. 2014 Novel heuristic density-based method for community detection in networks. *Physica A* **403**, 71–84. (doi:10.1016/j.physa.2014.01.043)
17. Liu W, Pellegrini M, Wang X. 2014 Detecting communities based on network topology. *Sci. Rep.* **4**, 5739. (doi:10.1038/srep05739)
18. Yang L, Jin D, Wang X, Cao X. 2015 Active link selection for efficient semi-supervised community detection. *Sci. Rep.* **5**, 9039. (doi:10.1038/srep09039)
19. Freeman LC. 1977 A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41. (doi:10.2307/3033543)

20. Brandes U. 2008 On variants of shortest-path betweenness centrality and their generic computation. *Soc. Netw.* **30**, 136–145. (doi:10.1016/j.socnet.2007.11.001)
21. Jaccard P. 1912 The distribution of the flora in the alpine zone. *New Phytol.* **11**, 37–50. (doi:10.1111/j.1469-8137.1912.tb05611.x)
22. Goldberg DS, Roth FP. 2003 Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA* **100**, 4372–4376. (doi:10.1073/pnas.0735871100)
23. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. 2002 Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555. (doi:10.1126/science.1073374)
24. Sibson R. 1973 SLINK: an optimally efficient algorithm for the single-link cluster method. *Comp. J.* **16**, 30–34. (doi:10.1093/comjnl/16.1.30)
25. Defays D. 1977 An efficient algorithm for a complete link method. *Comp. J.* **20**, 364–366. (doi:10.1093/comjnl/20.4.364)
26. Schütze H, Silverstein C. 1997 Projections for efficient document clustering. *ACM SIGIR Forum* **31**, 74–81. (doi:10.1145/278459.258539)
27. Zachary W. 1977 An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473.
28. The *C. elegans* Sequencing Consortium. 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018. (doi:10.1126/science.282.5396.2012)
29. Nerima B, Nilsson D, Mäser P. 2010 Comparative genomics of metabolic networks of free-living and parasitic eukaryotes. *BMC Genomics* **11**, 217. (doi:10.1186/1471-2164-11-217)
30. Newman ME, Strogatz SH, Watts DJ. 2001 Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118. (doi:10.1103/PhysRevE.64.026118)
31. Zhou H. 2003 Distance, dissimilarity index, and network community structure. *Phys. Rev. E* **67**, 061901. (doi:10.1103/PhysRevE.67.061901)
32. Aldecoa R, Marín I. 2011 Deciphering network community structure by surprise. *PLoS ONE* **6**, e24195. (doi:10.1371/journal.pone.0024195)