An *a posteriori* measure of network modularity

Timothée Poisot[1,2,*]

[1] *Département de biologie, chimie et géographie, Université du Québec à Rimouski, 300 Allée des Ursulines, Rimouski (QC), G5L 3A1, Canada*
[2] *Québec Centre for Biodiversity Sciences, Stewart Biological Sciences Building, 1205 Dr. Penfield Avenue, Montréal (QC), H3A 1B1, Canada*
[*] *email:* `timothee.poisot@uqar.ca`

Abstract

Measuring the modularity of networks, and how it deviates from random expectations, important to understand their structure and emerging properties. Several measures exist to assess modularity, which when applied to the same network, can return both different modularity values (*i.e.* different estimates of how modular the network is) and different module compositions (*i.e.* different groups of species forming said modules). More importantly, as each optimization method uses a different optimization criterion, there is a need to have an *a posteriori* measure serving as an equivalent of a goodness-of-fit. In this article, I propose such a measure of modularity, which is simply defined as the ratio of interactions established between members of the same modules *vs.* members of different modules. I apply this measure to a large dataset of 290 ecological networks representing host–parasite (bipartite) and predator–prey (unipartite) interactions, to show how the results are easy to interpret and present especially to a broad audience not familiar with modularity analyses, but still can reveal new features about modularity and the ways to measure it.

1 Introduction

Modularity, the fact that groups of nodes within a network interact more frequently with themselves than with other nodes, is an important property of several systems, including genetic [1, 2], informatics [3], ecological [4], and socio-economic [5] interactions, as well as biogeographic patterns [6, 7] and disease spread management [8]. Because of the relevance of modularity for network properties, it is important to assess it correctly. There exists several methods to measure network modularity, some of which rely on the optimization of a given criterion [9, 10], label propagation [11], or combination of these approaches [12, 13]. These methods return two elements. The first is a value of modularity for the networks, most often within the $0-1$ interval. Each method often has a threshold value, above which a network is considered to be modular. Increasing values reflect and increasingly modular structure. The second element is a "community partition", *i.e.* the attribution of each node to a module.

Recently, Thébault [7] showed that different measures of modularity tailored to presence/absence matrices (*i.e.* networks in which links have no weight), gave roughly equal estimates of the significance of modularity, but differed in the community partition they returned (*i.e.* the identity of nodes composing each module varied). In such situations, one might look for a way to choose which community partition should be used. The challenge is this situation is that the citeria used by each optimisation method cannot be meaningfully compared, and so there is a need for *a posteriori* measurement of how strong the modular structure is, regardless of the method used to obtain the community partition. More importantly, this criterion should be *different* than the one used to track the progress of any optimisation algorithm.

An important feature of modular networks is the occurrence of interactions between nodes of different modules. They contribute to the propagation of disturbances

[4], flow of information [14, 15], and cross- regulation of biological processes [16], *inter alia* [17]. In addition to measuring how modular the network is, determining to what extent modules are connected, and to identify nodes and edges responsible for connecting modules, is thus valuable information. In this article, I propose an *a posteriori* measure of the proportion of interactions established between modules, *i.e.* edges connecting different communities. I apply this measure to the community partition identified by the Louvain method on 290 ecological networks, and show that it behaves in a similar way to other modularity measures.

2 The measure

In this contribution I define the *realized modularity*, termed Q_R. Q_R measures the extent to which edges, within a network, are established between nodes belonging to the same module. For E edges in a network, if W of them are established between members of the same module, then

$$Q_R = \frac{W}{E}. \tag{1}$$

When there are no between-module links, then $W = E$ and Q_R takes the maximal value of 1. When between-module interactions are as numerous as within-modules interactions, then $W = E/2$, and Q_R takes the minimal value of $1/2$. To express the *realized modularity* as a value between 0 and 1, it is expressed as:

$$Q'_R = 2 \times Q_R - 1. \tag{2}$$

Note that Q'_R will yield values in the $[0;1]$ interval only if there are more edges established *within* than *between* modules. Although if modules are determined at random, Q'_R values are expected to be centered on 0, it is expected that they will increase when modules are properly optimized (only as far as the network is modular).

The main advantage of Q_R is that it is agnostic with regard to the measure used to optimize modularity (and even to the method by which the nodes were assigned to modules, which can be arbitrary), as it acts *a posteriori*, *i.e.* after nodes have been attributed to modules. Nonetheless, it assumes a simple yet functional definition of modularity: the fact that nodes interact more within than between modules. Given that measuring to which extent this is true, it can therefore be used to select the community detection method maximizing modularity. This measure works on most type of networks, as it makes no difference if links are directional, or if the networks are bipartite/unipartite. An illustration of this measure is given in Figure 1. This measure is purposefully simple, (i) so that it makes only minimal assumptions about what modularity is (except for the fact that in a modular network, nodes interact more within than between modules), or how it should be optimized, and (ii) because it is not meant to be used to optimize modularity, but to either compare the outcome of different methods, or present the value of modularity in a way that is straigthforward to interpretate.

A python implementation of this measure, using the `networkx` package, is proposed at `https://gist.github.com/tpoisot/4947006`. It reads data in the edge list format, and offers additional functions to generate null networks, as detailed in the following section.

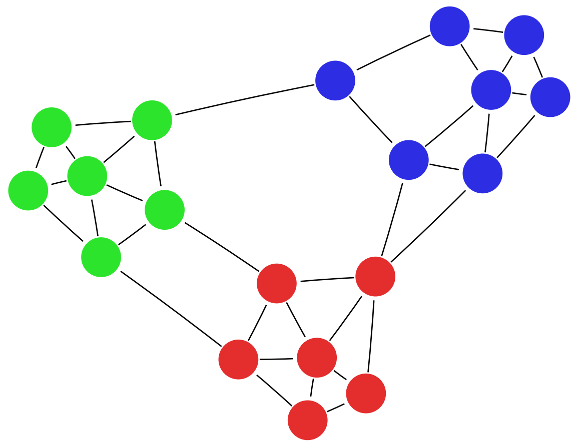

Figure 1: A cartoon depiction of a modular network with links between modules. Nodes of the same modules are identified by different colors. This network has a modularity (Louvain method) of $Q = 0.527$. Out of the 36 interactions, 31 are established within modules, and 5 between modules. This gives a Q_R value of 0.86, and $Q'_R = 0.72$.

3 Example application: realized modularity in ecological networks

In this section, I analyze the modular structure of a large dataset of 290 ecological networks (187 food webs and 113 host-parasite networks) published in previous meta-analyses [18, 19]. Modularity is an important feature of ecological interaction networks, which is linked to their resilience [20, 21], stability [7], biogeographic structure [22], functioning [23], and to the evolutionary mechanisms involved in their assembly [24]. Notably, the occurrence of interactions between and within modules plays a central role in the structure of pollination networks [4], and help buffer the effect of species extinctions [21].

The existence of interactions in ecological systems involves a large family of processes, ranging from abudance related [25, 26] (abundant species are more likely to interact together) to trait related [27] (pollination depends on the flower and insect having compatible morphologies, predators are constrained by the body-size of their preys). The interaction within these different families of mechanisms will drive heterogeneity in interaction strength [28]. Yet, the analysis of binary matrices (is there an interaction between a pair of species, or not), still has relevance to identify properties that are conserved across systems [29], especially given that one could argue that quantitative information on interaction strength is an additional level of information. The systems analyzed in this section are represented by their adjacency matrix, describing the presence or absence of an interaction. Bipartite networks have further been transformed into unipartite networks before analysis.

3.1 Data and analysis

I used the Louvain method [30] to detect modules, due to its rapidity and efficiency on large networks. The Louvain method works in two steps: first it optimizes modularity *locally*, through clustering of neighboring nodes. These clusters are, in the second steps, aggregated together, until modularity ceases to increase. This method is known to give values of modularity comparable to what is found using *e.g.* simulated annealing, and has been observed to give modules that have a functional relevance [30]. Once the partition is returned by the Louvain method, I recorded its realized modularity Q'_R, and its modularity Q (using the Newman and Girvan [31] measure).

For each network, I compared the values of Q and Q'_R on the empirical networks to their random estimate using a network null model. Because random networks will by chance (here meaning, as expected by networks having a given connectance and thus degree distribution, Poisot and Gravel [32]) display a modular (among other) structure, it is important to confront the empirical measures of both Q and Q'_R to their random expectations. The null model is defined as follows. For each node n of the network, I measured its degree d_n, its number of successors(the number of node it links to, or generality in ecological terms, as *per* [33]) g_n, and its number of predecessors (the number of nodes that link to it, or vulnerability) v_n. In each random network, for each pair of nodes (i,j), the probability that i interacts with j is given by

$$P(i \rightarrow j) = \frac{1}{2}\left(\frac{g_i}{d_i} + \frac{v_j}{d_j}\right), \tag{3}$$

and conversely for $P(j \rightarrow i)$. This null model allowed the generation of pseudo-random networks through a Bernoulli process (in each replicate, the occurrence of a link is randomly determined), with the same expected connectance, and the same expected distribution of degrees, generality, and vulnerability, as the original one (these properties are also conserved at the *node* level). For each of the 290 networks, 1000 pseudo-random replicates are generated. For each of them, the average value of Q_R and Q'_R are estimated along with their 90% confidence interval. When the empirical value lies outside the confidence interval, it can be assumed that the modular structure of the network is different than expected by chance.

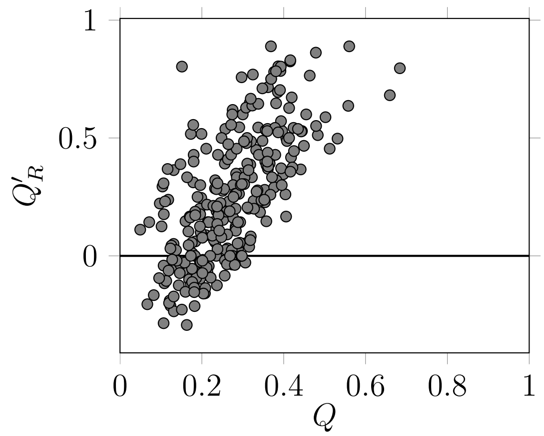

Figure 2: Relationship between the modularity of the best partition using the Louvain method and the *a posteriori* realized modularity. There exists a strong, positive relationship between the two variables. Worth noting is that fact that, for some networks, the best partition resulted in *negative* versions of Q'_R, *i.e.* there were more interactions between than within modules. Each dot correspond to a network.

3.2 Results and Discussion

There is a strong, positive relationship, between the values of Q'_R and the values of Q (Pearsons's product-moment correlation coefficient, as implemented in `R 2.15` [34], $\rho = 0.64$, 288 d.f., $p < 10^{-6}$), *i.e.* networks for which a high modularity is detected tend to have relatively few between-module links (Figure 2). It is worth noting that some Q'_R values were negative: in some cases, the best community division resulted in more interactions between than within modules. This result highlights why using an *a posteriori* measure is useful: other measures of modularity do not reveal the fact that there were more interactions between than within modules. In the dataset examined, most of the networks with a modularity lower than 0.2 had a negative realized modularity. This result suggests that discussing the modularity of such

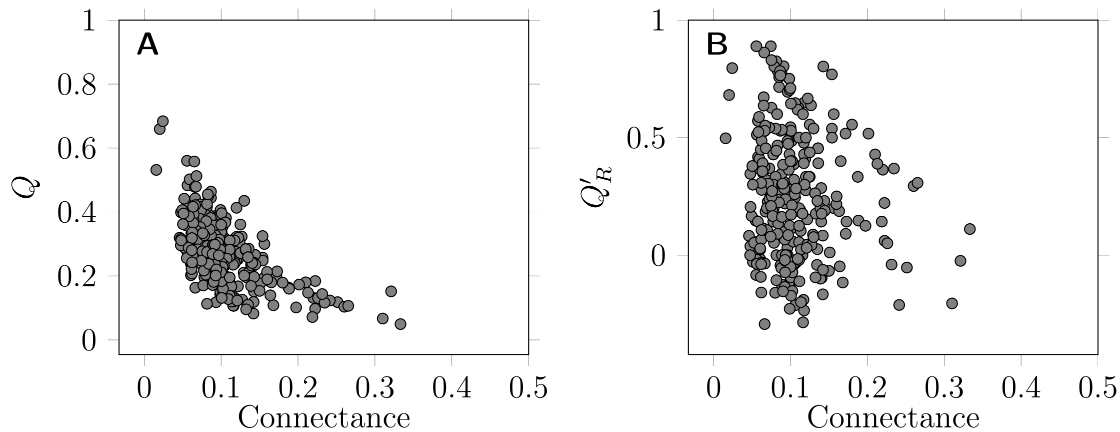

Figure 3: Relationship between the two measures of modularity and network connectance. **A.** Q is negatively affected by connectance, *i.e.* densely connected networks are more likely not to be modular. **B.** Q'_R is not affected by connectance, allowing to use it to compare different networks. Each dot correspond to a network.

networks makes little sense, as their modules are not more densely connected, within a module, than other random collections of nodes within the graph. Q and Q'_R have different relationships with connectance (Figure 3). Increased connectance values resulted in lower modularity ($\rho = -0.61$, 288 d.f., $p < 10^{-6}$), but had no impact on Q'_R. This is a desirable property, as it allows easy comparison with the Q'_R values of networks with extremely different connectances.

There is a linear relationship between the deviation from random expectation of Q and Q'_R ($\rho = 0.78$, 288 d.f., $p < 10^{-6}$ – Figure 4). The deviations (respectively ΔQ and $\Delta Q'_R$) are calculated as the empirical value, minus the average of the values on the networks generated by the null model. As an example, a ΔQ less than zero indicates that the empirical network is less modular than expected by chance. Confidence intervals for the average of the null models were typically very narrow (not represented in the figure to avoid cluttering – see associated original dataset), probably owing to the fact that the null model is restrictive on the type of networks which are generated. It is worth noting that for some networks, the diagnostic of the null model analysis conflicted. In a vast majority of the situations, this corresponds to networks having a lower modularity than expected by chance, yet having a higher realized modularity (dots in the upper left corner of Fig. 4). In this type of situations, whereas one would usually conclude that the networks are not significantly modular, the identified modules are nonetheless more densely connected (internally) than they are with the rest of the network. Because the dataset presents these contrasted situations, it allows to understand how the measure reacts to different network structures. Depending on whether the true modularity, or the realized modularity, is the most relevant metric of the processes studied, the interpretation of the null models for these networks will be different.

Finally, for the unipartite network dataset, I compare the results of three alternative methods of community detection (the walktrap, spinglass, and edge betweeness methods, as implemented in the *igraph* library). For each of the unipartite networks, I computed the value of Barber's Q, and Q'_R, on the best partition found. The strong correlation between Q and Q'_R were observed for the spinglass method ($\rho =$ 0.61,165 d.f., t $=$ 10.02), and the weakest for the edge-betweenness method ($\rho = 0.04$, non-significant at $\alpha = 0.05$). The walktrap algorithm gave results in between ($\rho =$ 0.489, 165 d.f., t $=$ 7.20). For both the walktrap and edge-betweeness methods, several networks had negative values of Q'_R, which indicates that the "best" community partition had more links between than within modules. The spinglass method had,

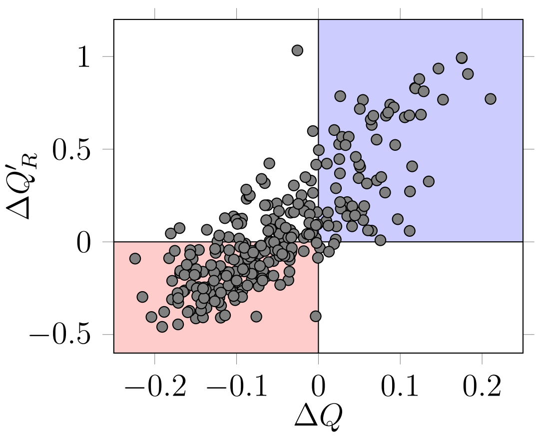

Figure 4: Linear relationship between the deviation from random expectation in Q and Q'_R. Networks in the red area are detected as being less modular than expected both by Q'_R and Q, while networks in the blue area are detected as being more modular. Although the agreement between the two measures is good (see main text for statistics), some networks are detected as having a higher than expected realized modularity Q'_R, despite a lower than expected modularity Q. Each dot correspond to a network.

by contrast, less than 8% of all networks with values of Q'_R lower than 0, meaning that this algorithm should be prefered when one wants to group nodes in densely connected clusters. This result reinforces the statement made by Thébault [7], *i.e.* that several modularity optimisation methods will return best modular structure that widely in their properties; thus, there is a need for *a posteriori* comparison of these outputs.

4 Conclusions

The Q'_R measure presented here allows the estimation of the proportion of interactions established between different modules in a network. This measure can be analyzed much in the same way as other measures of modularity, but is applied *a posteriori*. As such, it can help choose the "best" community partition according to the property of the network that one wants to maximize. For example, choosing the partition giving the lowest Q'_R can help identify which species are more likely to act as connectors between different modules. Ultimately, this information may have some practical relevance as a decision tool. Saavedra et al. [5] showed that different nodes contribute differently to overall network properties. In a context in which networks are increasingly being used as management tools to adress *e.g.* conservation or pest management [8], knowing the realized modularity, and developping methods to estimate which species have the highest impact on it, can allow the design of efficient policies to maximize, or decrease, the ability of network modules to interact.

Competing interests

No competing interests were disclosed.

Grant information

TP is funded by a PBEEE-FQRNT post-doctoral scholarship, and thanks the EEC Canada Research Chair for providing computational support.

Acknowledgements: Thanks are due to the maintainers and contributors of the free `networkx`, `scipy`, `igraph` and `numpy` packages used in this project, and to Scott Chamberlain for discussions.

References

[1] Carlos Espinosa-Soto and Andreas Wagner. "Specialization can drive the evolution of modularity". In: *PLoS Computational Biology* 6.3 (2010), e1000719.

[2] A. Bauer-Mehren et al. "Gene-Disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases". In: *PLoS One* 6.6 (2011), e20284.

[3] M. A. Fortuna, J. A. Bonachela, and S. A. Levin. "Evolution of a modular software network". In: *Proceedings of the National Academy of Sciences of the United States of America* 108.50 (Nov. 2011), 19985–19989.

[4] Jens M Olesen et al. “The modularity of pollination networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.50 (2007), 19891–19896.

[5] Serguei Saavedra et al. “Strong contributors to network persistence are the most vulnerable to extinction”. In: *Nature* 478.7368 (Sept. 2011), 233–235.

[6] Daniel W. Carstensen et al. “Biogeographical modules and island roles: a comparison of Wallacea and the West Indies”. In: *Journal of Biogeography* (Nov. 2011), no–no.

[7] Elisa Thébault. “Identifying compartments in presence-absence matrices and bipartite networks: insights into modularity measures”. In: *Journal of Biogeography* (Nov. 2012). Ed. by Joseph Veech, n/a–n/a.

[8] Iadine Chadès et al. “General rules for managing and surveying networks of pests, diseases, and endangered species.” In: *Proceedings of the National Academy of Sciences of the United States of America* 108.20 (May 2011), 8323–8328.

[9] Mark E J Newman. “Modularity and community structure in networks.” In: *Proceedings of the National Academy of Sciences of the United States of America* 103.23 (June 2006), 8577–82.

[10] Xiang-Sun Zhang and Rui-Sheng Wang. “Optimization analysis of modularity measures for network community detection”. In: *The Second International Symposium on Optimization and Systems Biology*. Lijiang, China, 2008, 13–20.

[11] Michael Barber. “Modularity and community detection in bipartite networks”. In: *Physical Review E* 76.6 (Dec. 2007).

[12] Xin Liu and Tsuyoshi Murata. “Community detection in large-scale bipartite networks”. In: *Transactions of the Japanese Society for Artificial Intelligence*. Vol. 5. 2010, 184–192.

[13] Flavia Maria Darcie Marquitti et al. *MODULAR: Software for the Autonomous Computation of Modularity in Large Network Sets*. arXiv e-print 1304.2917. Apr. 2013.

[14] Gio Wiederhold. “Mediators in the architecture of future information systems”. In: *The IEEE Computer Magazine* (1992), 1–36.

[15] Jure Leskovec et al. “Statistical properties of community structure in large social and information networks”. In: *Proceeding of the 17th international conference on World Wide Web - WWW ’08*. New York, New York, USA: ACM Press, 2008, p. 695.

[16] L H Hartwell et al. “From molecular to modular cell biology.” In: *Nature* 402.6761 Suppl (Dec. 1999), C47–52.

[17] Martin Rosvall and CT Bergstrom. “Maps of random walks on complex networks reveal community structure”. In: *Proceedings of the National \ldots* 105.4 (Jan. 2008), 1118–23.

[18] Dominique Gravel et al. “Trophic theory of island biogeography.” In: *Ecology Letters* 14.10 (Oct. 2011), 1010–1016.

[19] Timothée Poisot et al. “The dissimilarity of species interaction networks.” In: *Ecology Letters* 15.12 (Sept. 2012), 1353–1361.

[20] Miguel Angel Fortuna et al. “Nestedness versus modularity in ecological networks: two sides of the same coin?” In: *Journal of Animal Ecology* 78 (2010), 811–817.

[21] Daniel B Stouffer and Jordi Bascompte. “Compartmentalization increases food-web persistence.” In: *Proceedings of the National Academy of Sciences of the United States of America* 108.9 (Feb. 2011), 3648–3652.

[22] Cesar O Flores, Sergi Valverde, and Joshua S Weitz. “Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages”. In: *The ISME Journal* (Nov. 2012), 1–13.

[23] Elisa Thébault and Michel Loreau. “Food-web constraints on biodiversity–ecosystem functioning relationships”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.25 (2003), 14949–14954.

[24] Cesar O Flores et al. “Statistical structure of host-phage interactions.” In: *Proceedings of the National Academy of Sciences of the United States of America* 108.28 (June 2011), E288–297.

[25] Nico Blüthgen, Florian Menzel, and Nils Blüthgen. “Measuring specialization in species interaction networks.” In: *BMC ecology* 6 (Jan. 2006), p. 9.

[26] Elsa Canard et al. “Emergence of Structural Patterns in Neutral Trophic Networks”. In: *PLoS ONE* 7.8 (Aug. 2012), e38295.

[27] Ignasi Bartomeus. "Understanding Linkage Rules in Plant-Pollinator Networks by Using Hierarchical Models That Incorporate Pollinator Detectability and Plant Traits". In: *PLoS ONE* 8.7 (July 2013), e69200.

[28] E.L. Berlow et al. "Quantifying variation in the strengths of species interactions". In: *Ecology* 80.7 (1999), 2206–2224.

[29] Jennifer A Dunne. "The Network Structure of Food Webs". In: *Ecological networks: Linking structure and dynamics*. Ed. by Jennifer A Dunne and Mercedes Pascual. Oxford University Press, 2006, 27–86.

[30] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008.

[31] MEJ Newman and M Girvan. "Finding and evaluating community structure in networks". In: *Physical review E* (2004).

[32] Timothée Poisot and Dominique Gravel. "When is a network complex? Connectance drives degree distribution and emerging network properties". In: *PeerJ preprints* 1.e50v3 (2013).

[33] Thomas W Schoener. "Food webs from the small to the large". In: *Ecology* 70.6 (1989), 1559–1589.

[34] R. Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008.