



# MODULAR: software for the autonomous computation of modularity in large network sets

Flavia Maria Darcie Marquitti, Paulo Roberto Guimarães Jr, Mathias Mistretta Pires and Luiz Fernando Bittencourt

*F. M. D. Marquitti (flamarquitti@gmail.com), P. R. Guimarães Jr and M. M. Pires, Depto de Ecologia, Univ. de São Paulo, São Paulo, SP 05508-900, Brazil. – L. F. Bittencourt, Inst. de Computação, Univ. Estadual de Campinas, Av. Albert Einstein, 1251, Campinas, SP 13083-852, Brazil.*

Many ecological systems can be represented as networks of interactions. A key feature in these networks is their organization into modules, which are subsets of tightly connected elements. We introduce MODULAR to perform rapid and autonomous calculation of modularity in network sets. MODULAR reads a set of files representing unipartite or bipartite networks, and identifies modules using two different modularity metrics widely used in the ecological networks literature. To estimate modularity, the software offers five optimization methods to the user. The software also includes two null models commonly used in studies of ecological networks to verify how the degree of modularity differs from two distinct theoretical benchmarks.

Ecological systems can be seen as networks in which the elements, such as habitat patches within a landscape or species within communities, are represented by nodes and patch connectivity or species interactions are depicted as edges connecting the nodes (Pascual and Dunne 2006, Urban et al. 2009). The way such connections are organized affects system dynamics (Stouffer and Bascompte 2011) and, thus, how the system will respond to changes such as species loss (Dunne et al. 2002) or changes in the ecological connectivity among patches (Cumming et al. 2010). Several properties have been used to characterize the organization of networks, i.e. the network topology (Pascual and Dunne 2006, Carstensen et al. 2012). A recurrent pattern in ecological networks is modularity, which is also termed compartmentalization, clustering, or community structure (Boccaletti et al. 2006). Modules are cohesive groups of highly connected nodes that are loosely connected to other nodes in the network (Newman and Girvan 2004, Olesen et al. 2007).

Modular organization has been found in networks describing different ecological systems, such as resource use by animal populations (Araújo et al. 2008), food webs (Krause et al. 2003, Allesina and Pascual 2009), mutualistic interactions between plant species and their pollinators (Olesen et al. 2007), antagonistic interactions between parasites and their hosts and between plants and their herbivores (Fortuna et al. 2010, Pires and Guimarães 2013), and the spatial connectivity of metapopulations (Bodin and Norberg 2007, Dale and Fortin 2010). In these ecological networks, modularity will emerge if certain groups of individuals,

species, or habitat patches show more interactions among each other than with other groups within the network (Guimarães et al. 2007, Dale and Fortin 2010, Carstensen et al. 2012, Tinker et al. 2012). Because the degree of modularity measures how dense the connections between different groups of elements in an ecological system are, systems that largely differ in the degree of modularity often differ in their ecological and evolutionary dynamics (Guimarães et al. 2007, Olesen et al. 2007, Cumming et al. 2010, Thébault and Fontaine 2010, Hagen et al. 2012).

Motivated by the relevance of modularity for the dynamics of ecological systems, the detection of modularity became an important aspect of studies that analyze the organization of ecological networks (Cumming et al. 2010). Modularity can be measured using different indices. One of the most popular indices is the  $Q$  metric (also known as  $M$ ) (Newman and Girvan 2004, Guimerà and Amaral 2005, Olesen et al. 2007, Fortuna et al. 2010). For a given partition of a network into modules,  $Q$  is given by the difference between the observed fraction of edges connecting nodes in the same module and the expected fraction of edges connecting nodes in the same module if connections were to occur at random. For any given network, the researcher is often interested in the partition that actually best describes the presence of modules, i.e., the partition that maximizes  $Q$ . Because finding the maximal modularity (maximal  $Q$ ) in a network is an NP-hard problem (Ruan and Zhang 2007), there is no known algorithm to find the maximum  $Q$  in polynomial time. It is thus necessary to use optimization approaches that cannot guarantee the maximization of modularity,

but can provide an approximate solution in reasonable running time. Heuristic algorithms can thus be used to find a network partitioning with approximately the largest number of edges within the modules and the lowest number of edges between modules (see details of the metrics in Supplementary material Appendix 1). Although module identification is scale-dependent, optimization algorithms can be used to test module consistency across multiple scales, testing the effects of resolution on module detection (Fortunato and Barthélemy 2007).

Finding the partition with the highest modularity in a large network is often time consuming. Moreover, the analysis of large sets of data and the subsequent testing of the results against theoretical benchmarks that are represented by ensembles formed by thousands of replicates is a common procedure in biology (Gotelli and Graves 1996, Gotelli 2001). In this sense, a major constraint in the analysis of the modularity of ecological networks is the lack of programs that allow fast and autonomous computation of modularity for researchers who are not familiar with programming. Here, we introduce the software MODULAR for the computation of the modularity and the identification of modules in multiple complex networks. Many algorithms have been proposed for finding the partition that maximizes the value of  $Q$ , and some of these are publicly available. However, to the best of our knowledge, there is no software available that allows multiple uses of different metrics and optimization algorithms in a user-friendly way that accelerates the workflow of the ecologist. MODULAR was developed to allow the user to automatically compute the modularity of several input files and to allow the user to choose the optimization algorithm that best matches the user's needs.

## MODULAR features

MODULAR was developed in the *C* language and uses features from the *igraph-0.6* library (Csárdi and Nepusz 2006) and the GNU Scientific Library (GSL) (Gough 2009) (see details in Supplementary material Appendix 2). MODULAR was designed to facilitate and accelerate the detection of modules in multiple networks through maximization of the degree of modularity. To achieve this task, the maximization of modularity is automatically performed for a set of input files containing representations of bipartite networks, such as those depicting species occurrence across islands (Carstensen et al. 2012), or unipartite networks, such as spatial networks describing habitat connectivity for a given species (Dale and Fortin 2010).

When running MODULAR with unipartite networks, only the  $Q$  metric is available (Newman and Girvan 2004). If the input data files represent bipartite networks, the user can choose between two different modularity metrics: Newman and Girvan's –  $Q$  (Newman and Girvan 2004) or Barber's modularity –  $Q_B$  (Barber 2007), which is a modification of the  $Q$  metric for bipartite networks. Bipartite networks have two sets of nodes, and interactions can only occur between nodes of different sets (Costa et al. 2007).  $Q_B$  differs from  $Q$  in that in  $Q_B$ , the second term of the

equation, which measures the expected fraction of edges within modules if edges are placed randomly, only considers interactions among nodes of different sets (see details of metrics in Supplementary material Appendix 1). Thus, researchers interested in measuring the modularity of unipartite networks should use  $Q$ , whereas those interested in the modularity of bipartite networks should use  $Q_B$ . Nevertheless, the two options,  $Q$  and  $Q_B$ , are available for bipartite networks because unipartite indexes have also been used in the ecological literature for bipartite networks (Olesen et al. 2007, Fortuna et al. 2010, Carstensen et al. 2012). Although bipartite networks analyzed using  $Q$  and  $Q_B$  show similar values of modularity, their module definitions differ strongly (Thébaud 2013). Thus, the user could explore how different definitions of modularity would affect their results and conclusions by exploring network organization with distinct metrics.

If the user chooses the traditional  $Q$  metric, there are five optimization algorithms that can be used to perform the search for the partition of the network into modules that maximizes the modularity index: 1) fast greedy (FG) (Clauset et al. 2004, Wakita and Tsurumi 2007), 2) simulated annealing (SA) (Guimerà and Amaral 2005), 3) spectral partitioning (SP) (Newman 2006), 4) a hybrid of simulated annealing and spectral partitioning (Hyb-SP), and 5) a hybrid of simulated annealing and fast greedy (Hyb-FG). The optimization algorithms differ in the method used to search for the network partition that maximizes the modularity measurement (see details of MODULAR functioning and optimization algorithms in Supplementary material Appendix 3). Because the running time can vary considerably, the choice of the optimization algorithm becomes particularly important. We tested the performance of optimization algorithms for bipartite and unipartite ecological networks. The different algorithms represent a trade-off between speed and the quality of the estimate of the maximal modularity. Hybrid algorithms are faster and present higher values of modularity than the SA algorithm alone. The other two algorithms, FG and SP, are faster than all the algorithms that include the SA process, and they seem to reach similar results to the SA-based algorithms for small networks. Thus, for small networks or for exploratory initial analyses of modularity, we suggest the user should use FG and SP, whereas SA and hybrid SA-based algorithms should be used for in-depth analysis of large networks. We present the results for a set of ecological networks in the Supplementary material Appendix 4.

To verify whether the modularity found by the maximization of each metric significantly differs from a theoretical benchmark, the user has the option of running two different null models and specifying how many replicates each null model will generate. We included unipartite and bipartite versions of two of the most common null models that are used in studies of ecological networks: 1) the Erdős-Rényi model (Erdős and Rényi 1959) and 2) the 'null model 2' (Bascompte et al. 2003). The first null model generates networks of the same size and connectance with edges randomly distributed among the nodes. The second null model also considers a similar heterogeneity of edges among nodes as observed in the original network, maintaining part of the structure of the network by approximately

conserving the original distribution of edges per node. Thus, in a nutshell, the use of the two models may allow the user to infer if the observed degree of modularity could be reproduced by random networks with similar node richness and number of edges (i.e. if the Erdős-Rényi model reproduces the observed modularity), or if the user also needs heterogeneity in number of interactions (i.e. only Bascompte's null model reproduces the modularity), or if the user needs additional factors not incorporated in either null model to reproduce the observed degree of modularity (i.e. both null models fail in reproducing the modularity of the real network); see details of the null models in Supplementary material Appendix 5. However, null model analysis is not restricted to these two null models in MODULAR. Because MODULAR can utilize large sets of input data, the user can also test the modularity of the networks generated by other null models by using those as input data.

MODULAR is an open source software program licensed under the GNU General Public License version 3. It can be downloaded from <http://sourceforge.net/projects/programmodular/>. In the future, we are planning to add new algorithms and optimization methods for the calculation of modularity. In addition, new metrics that analyze modularity at the node level can be added to MODULAR.

To cite MODULAR or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 0':

Marquitti, F. M. D., Guimarães Jr, P. R., Pires, M. M. and Bittencourt, L. F. 2014. MODULAR: software for the autonomous computation of modularity in large network sets. – *Ecography* 37: 000–000 (ver. 0).

*Acknowledgements* – We thank the igraph developers for promptly solving all of the reported issues in the external interactions of the library. We would also like to thank all of our colleagues who encouraged us to persist in developing this software and who tested it in early stages. We are very thankful to Pedro Jordano for his comments and suggestions. FMDM; MMP; and PRG are funded by the São Paulo Research Foundation (FAPESP): grants no. 2011/11346-0 and no. 2012/05935-5; no. 2009/54567-6; and no. 2009/54422-8.

## References

Allesina, S. and Pascual, M. 2009. Food web models: a plea for groups. – *Ecol. Lett.* 12: 652–662.

Araújo, M. S. et al. 2008. Network analysis reveals contrasting effects of intraspecific competition on individual vs. population diets. – *Ecology* 89: 1981–1993.

Barber, M. J. 2007. Modularity and community detection in bipartite networks. – *Phys. Rev. E* 76: 066102.

Bascompte, J. et al. 2003. The nested assembly of plant–animal mutualistic networks. – *Proc. Natl Acad. Sci. USA* 100: 9383–9387.

Boccaletti, S. et al. 2006. Complex networks: structure and dynamics. – *Phys. Rep.* 424: 175–308.

Bodin, Ö. and Norberg, J. 2007. A network approach for analyzing spatially structured populations in fragmented landscape. – *Landscape Ecol.* 22: 31–44.

Carstensen, D. W. et al. 2012. Biogeographical modules and island roles: a comparison of wallacea and the west indies. – *J. Biogeogr.* 39: 739–749.

Clauset, A. et al. 2004. Finding community structure in very large networks. – *Phys. Rev. E* 70: 066111.

Costa, L. d. F. et al. 2007. Characterization of complex networks: a survey of measurements. – *Adv. Phys.* 56: 167–242.

Csárdi, G. and Nepusz, T. 2006. The igraph software package for complex network research. – *InterJournal Complex Systems*, 1695.

Cumming, G. S. et al. 2010. Network analysis in conservation biogeography: challenges and opportunities. – *Divers. Distrib.* 16: 414–425.

Dale, M. and Fortin, M.-J. 2010. From graphs to spatial graphs. – *Annu. Rev. Ecol. Evol. Syst.* 41: 21–38.

Dunne, J. A. et al. 2002. Network structure and biodiversity loss in food webs: robustness increases with connectance. – *Ecol. Lett.* 5: 558–567.

Erdős, P. and Rényi, A. 1959. On random graphs. – *Publ. Math. Debrecen* 6: 290–297.

Fortuna, M. A. et al. 2010. Nestedness versus modularity in ecological networks: two sides of the same coin? – *J. Anim. Ecol.* 79: 811–817.

Fortunato, S. and Barthélemy, M. 2007. Resolution limit in community detection. – *Proc. Natl Acad. Sci. USA* 104: 36–41.

Gotelli, N. J. 2001. Research frontiers in null model analysis. – *Global Ecol. Biogeogr.* 10: 337–343.

Gotelli, N. J. and Graves, G. R. 1996. Null models in ecology. – *Smithsonian Inst. Press*.

Gough, B. 2009. GNU scientific library reference manual, 3rd ed. – *Network Theory*.

Guimarães, P. R. et al. 2007. Interaction intimacy affects structure and coevolutionary dynamics in mutualistic networks. – *Curr. Biol.* 17: 1797–1803.

Guimerà, R. and Amaral, L. A. N. 2005. Functional cartography of complex metabolic networks. – *Nature* 433: 895–900.

Hagen, M. et al. 2012. Biodiversity, species interactions and ecological networks in a fragmented world. – *Adv. Ecol. Res.* 46: 89–210.

Krause, A. E. et al. 2003. Compartments revealed in food-web structure. – *Nature* 426: 282–285.

Newman, M. E. J. 2006. Finding community structure in networks using the eigenvectors of matrices. – *Phys. Rev. E* 74: 036104.

Newman, M. E. J. and Girvan, M. 2004. Finding and evaluating community structure in networks. – *Phys. Rev. E* 69: 026113.

Olesen, J. M. et al. 2007. The modularity of pollination networks. – *Proc. Natl Acad. Sci. USA* 104: 19891–19896.

Pascual, M. and Dunne, J. A. 2006. Ecological networks: linking structure to dynamics in food webs. – *Oxford Univ. Press*.

Pires, M. M. and Guimarães, P. R. 2013. Interaction intimacy organizes networks of antagonistic interactions in different ways. – *J. R. Soc. Interface* 10: 20120649.

Ruan, J. and Zhang, W. 2007. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. – In: *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on Ieee*, pp. 643–648.

Stouffer, D. B. and Bascompte, J. 2011. Compartmentalization increases food-web persistence. – *Proc. Natl Acad. Sci. USA* 108: 3648–3652.

Thébault, E. 2013. Identifying compartments in presence–absence matrices and bipartite networks: insights into modularity measures. – *J. Biogeogr.* 40: 759–768.

- Thébault, E. and Fontaine, C. 2010. Stability of ecological communities and the architecture of mutualistic and trophic networks. – *Science* 329: 853–856.
- Tinker, M. T. et al. 2012. Structure and mechanism of diet specialisation: testing models of individual variation in resource use with sea otters. – *Ecol. Lett.* 15: 475–483.
- Urban, D. L. et al. 2009. Graph models of habitat mosaics. – *Ecol. Lett.* 12: 260–273.
- Wakita, K. and Tsurumi, T. 2007. Finding community structure in mega-scale social networks. – In: *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 1275–1276.

Supplementary material (Appendix ECOG-00506 at <[www.oikosoffice.lu.se/appendix](http://www.oikosoffice.lu.se/appendix)>). Appendix 1–5.