Network modularity is widely misused in ecological analyses

Michael D. Catchen[1,2,3]

Abstract Stop using modularity maximization.

1 Introduction

Ecosystems are composed of interactions between species and their environment. These interactions form networks that enable the persistence of species, ecosystems, and the services ecosystems provide people. In the last few decades, *network science* has developed to understand networks across a variety of domains. This field has developed numerous quantitative tools for describing network structure, which have seen increasing adoption in ecosystem science in the burgeoning subfield of network ecology (Delmas *et al.* 2019). One such property is *modularity* (denoted Q), which is a metric that describes "how well" nodes of a network can be grouped into *modules*, first introduced in Newman & Girvan (2004). Modularity has been widely adopted as a metric of interest in ecological networks, and in principle the grouping of species into modules could contain biologically meaningful information.

Unfortunately, the most popular method identifying modules in ecological networks is Modularity Maximization (MM), which has many well documented flaws for robustly identifying modules in networks (Fortunato & Barthélemy 2007; Good *et al.* 2010; Lancichinetti & Fortunato 2011; Peixoto 2021).

As an alternative, we suggest methods for community detection based on Stochastic Block Models (Karrer & Newman 2011; Peixoto 2014; Yen & Larremore 2020) for identifying modules in ecological networks. Although they have seen some use in ecological networks (Allesina & Pascual 2009; Gauzens *et al.* 2013; O'Connor *et al.* 2020), modularity is still predominantly used in network ecology. In a brief literature survey, we found MM methods overwhelmingly prevelent in the analysis of ecological networks. Here we cover what modularity maximization is, and why it doesn't work for identifying modules/groups in networks. We then provide a brief primer on stochastic block models.

2 What is modularity?

Consider an undirected network defined by an adjacency matrix $\mathbf{A}$, where $A_{ij} = 1$ if nodes i and j share an edge, and 0 otherwise. Let $m = \sum_{i,j} A_{ij}$ denote the total number of edges in the network, and k_i be the degree (the number of edges) associated with node i. Let b_i denote the *group* (or module) that node i belongs to. Modularity (Q) is then defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(b_i, b_j)$$

where δ is a function that equals 1 if $b_i = b_j$, and equals 0 otherwise. It is essential to emphasize that **modularity is not a property of a network *alone***. It is only defined for a network *and a set of group assignments for each node*, $\vec{b}$.

This value can be interpreted intuitively as how many more edges exist between members of the same group than would be expected if edges were distributed "at random". As pointed out by Peixoto (2021), there is an implicit null model in what "at random" means in this definition, namely the Chung-Lu configuration model (Chung & Lu 2002), where the probability of an edge existing between nodes i and j is $\mathbb{E}[A_{ij}] = \frac{k_i k_j}{2m}$.

3 What is modularity maximization?

Modularity maximization (MM) is one of many potential methods for the problem of taking an observed network $\mathbf{A}$ and infering which group b_i each node i belongs to, and how many total groups $\mathcal{B}$ there are total (in network science literature, this problem is called *community detection*). MM originated during the mid-2000s (Newman & Girvan 2004) and was popularized through the efficeincy of the Clauset-Newman-Moore (CNM) algorithm (Clauset *et al.* 2004) and the Louvain algorithm (Blondel *et al.* 2008), both of which made implementation of MM feasible for very large networks (at the time, hundreds or thousands of nodes). Six years later after its proposal, Good *et al.* (2010) (with Clauset, architect of CNM, as senior author) showed that in practice communities identified via modularity maximization are fataly flawed for all but idealized networks, and advocated against its use in "in all but the most straightforward cases". More recently, Peixoto (2021) more thoroughly explores this issue, showing how MM can massively overfit and find highly modular partitions ($Q \approx 0.5$) in networks with no modular structure.

4 Why doesn't modularity maximization work?

As pointed out by Peixoto (2021), modularity maximization fails on two fronts: it simultaneously *overfits* (by finding clusters that have high modularity Q but are entirely

sporatic and unrelated to the mechanisms by which the network was generated) and *underfits* (by having a limit on the size of what communities are recoverable relative to the size of the whole network, called the *resolution limit* (Fortunato & Barthélemy 2007)).

4.1 Overfitting via a poor choice of objective function

The first issue with modularity maximization is the the modularity function Q has many local optima, with similar values of Q, but which correspond to qualitatively very different partitions $\vec{b}$. This was first reported in Good *et al.* (2010), who also show that Q_{max} is highly dependent on the number of clusters and the size of the network, and conclude—"[the] modules identified through modularity maximization should be treated with caution in all but the most straightforward cases" (Good *et al.* 2010).

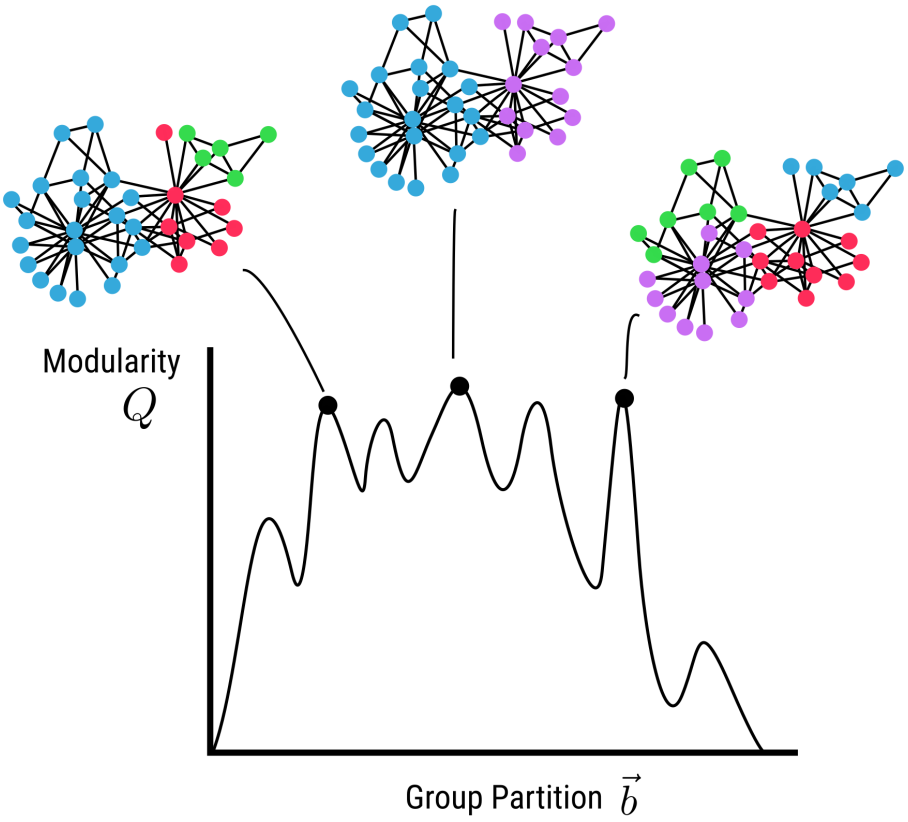

Figure 1: The issue with modularity maxmimization: there are many local optima with similar Q values that correspond to qualitatively very different group partitions.

4.2 Underfitting via the resolution limit

The second issue with modularity maximization is that is cannot identify communities that at smaller than a certain size. The threshold for smallest community identifiable via MM is a function of the total size over the network, and called the "resolution limit" in the network science literature (Fortunato & Barthélemy 2007; Lancichinetti & Fortunato 2011).

5 Modularity maximization is rampant in ecological network studies

We found in a survey of 50+ papers on ecological networks, modularity maximization is extremely common as the method for finding communities. The goal of this paper is not to shame or call-out specific papers, but to highlight that a widely adopted practice has fundemental flaws, and to advocate a principle alternative for community detection.

We suspect MM is so prolific because it is widely available in many packages for network analysis, including `bipartite`, which uses a method for modularity maximization for bipartite networks proposed by Dormann & Strauss (2014), and the very popular libraries `igraph` and `networkx`. Another widely applied method is from Guimerà & Nunes Amaral (2005), which uses simulated annealing for MM. The prolific availability of software to run MM-based community detection leads researchers down the "path of least resistance".

6 What instead of modularity maximization?

The state-of-the-art for community detection in networks are using a family of models called Stochastic Block Models (SBMs). Although the initial idea dates back several decades (Holland *et al.* 1983), modern research into using SBMs for community detection was spurred by regonition of the flaws with modularity maxmization (Good *et al.* 2010). SBMs have several advantages over modularity maximization. SBM inference is naturally posed as a Bayesian inference problem (Hofman & Wiggins 2008), which allows us to explicitly account for uncertainty in our estimate of the best node partition $\vec{b}$. Further, hierarchical SBMs (Peixoto 2014), where each block is itself an SBM, enables multi-scale community detection.

6.1 What is a stochastic block model?

SBMs are a *probabilistic generative model*. This means for a fixed set of input parameters, SBMs can be sampled to produce different possible realizations of networks from the *distribution* of possible networks given the input parameters. In their simplest form, SBMs take a partition of the nodes into a groups $\vec{b}$, and a mixing or block matrix $\mathbf{M}$, where $\mathbf{M}_{b_i,b_j}$ is the probablity of an edge existing between nodes in groups b_i and b_j respectively.

This enables much more flexability in the types of community structure exist in networks. Modularity maximization can only capture *one type* of community structure—*assortative* communities, where links within communities are more common that those between communities. In contrast, community structure in networks can take on a variety of different forms: assortative, disassortative (where *between group* edges are more likely than *within group*), core-periphery (where a set of densely connected nodes form a 'core', and other 'periphery' nodes that have few edges and tend to be attached to core nodes), and ordered (like trophic levels in a food-web).

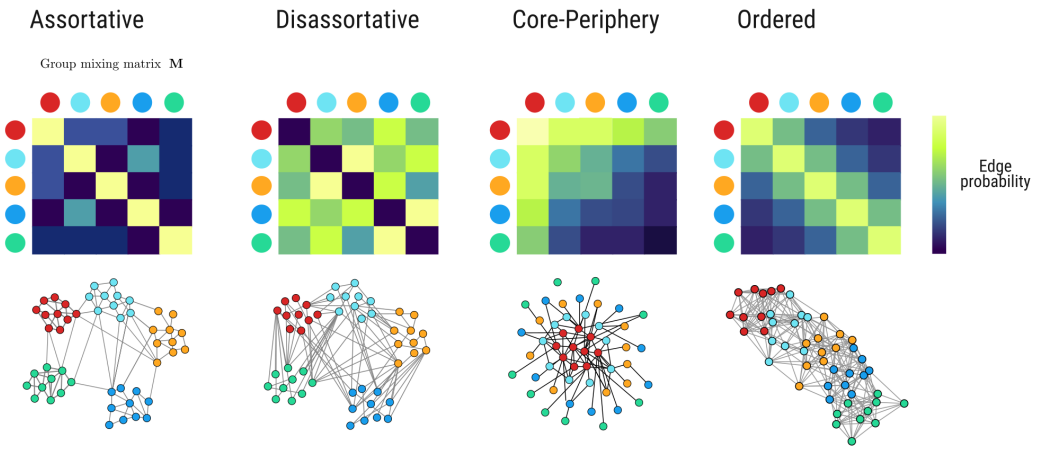

Figure 2: "Adapted from Clauset (2022). The mixing matrix $\mathbf{M}$ for different SBMs that account for different types of community structure."

6.2 How do we infer community structure from stochastic block models?

We can use Markov Chain Monte Carlo (MCMC) sampler to take an observed matrix A and obtain an estimate of the posterior distribution of the mixing matrix and group assignments, $P(\mathbf{M}, \vec{b}|\mathbf{A})$. To do this, we need to define the likelihood of observing some network $\mathbf{A}$ from a given community partition $\vec{b}$, and mixing matrix $\mathbf{M}$. There are differences in the best way to define both likelihood and priors depending on underlying assumptions about network structure.

For unipartite networks, a common version is the Degree-Corrected SBM (DC-SBM, Karrer & Newman 2011), which explicitly accounts for the degree distribution by including the empirical degree sequence in the likelihood of observing each graph.

Nested SBMs (Peixoto 2014). In NSBMs, each "block" $\mathbf{M}_{b_i b_j}$ is *itself* another SBM. This

enables multi-scale community detection that can circumvent the issue of resolution limits from modularity maximization.

Modern work on SBMs typically focuses on variants of the *microcanonical version* of both the DC-SBM and NSBM (Peixoto 2017). Here *microcanonical* is terminology being adopted from statistical mechanics, which in practice means these models are defined for a fixed degree sequence (number of edges per nodes). For a thorough recent-ish review of block modeling, see Lee & Wilkinson (2019).

Yen & Larremore (2020) develops a model specifically for bipartite networks, where the bipartite structure is directly incorporated into the likelihood, improving performance for detecting communities in bipartite networks over DC-SBM.

7 Conclusion

In summary, community detection is great, but modularity maximization is useless. There are times when modularity, as a method of quantifying the assortativity of edges in a graph given a set of group assignments $\vec{b}$, could correspond to an interesting ecological question. However, using modularity *as the criteria* to select the group assignments is too unreliable to be the basis ecological conclusions. As an alterative, we should use *stochastic block models* to infer the structure of modules within ecological networks.

References

Allesina, S. & Pascual, M. (2009). Food web models: A plea for groups. *Ecology Letters*, 12, 652–662.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008.

Chung, F. & Lu, L. (2002). Connected Components in Random Graphs with Given Expected Degree Sequences. *Annals of Combinatorics*, 6, 125–145.

Clauset, A. (2022). Lecture Notes – Network Analysis and Modeling.

Clauset, A., Newman, M.E.J. & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70, 066111.

Delmas, E., Besson, M., Brice, M.-H., Burkle, L.A., Riva, G.V.D., Fortin, M.-J., *et al.* (2019). Analysing ecological networks of species interactions. *Biological Reviews*, 94, 16–36.

Dormann, C.F. & Strauss, R. (2014). A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, 5, 90–98.

Fortunato, S. & Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 36–41.

Gauzens, B., Legendre, S., Lazzaro, X. & Lacroix, G. (2013). Food-web aggregation, methodological and functional issues. *Oikos*, 122, 1606–1615.

Good, B.H., de Montjoye, Y.-A. & Clauset, A. (2010). The performance of modularity maximization in practical contexts. *Physical Review E*, 81, 046106.

Guimerà, R. & Nunes Amaral, L.A. (2005). Functional cartography of complex metabolic networks. *Nature*, 433, 895–900.

Hofman, J.M. & Wiggins, C.H. (2008). Bayesian Approach to Network Modularity. *Physical Review Letters*, 100, 258701.

Holland, P.W., Laskey, K.B. & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5, 109–137.

Karrer, B. & Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83, 016107.

Lancichinetti, A. & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E*, 84, 066122.

Lee, C. & Wilkinson, D.J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4, 1–50.

Newman, M.E.J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69, 026113.

O'Connor, L.M.J., Pollock, L.J., Braga, J., Ficetola, G.F., Maiorano, L., Martinez-Almoyna, C., *et al.* (2020). Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. *Journal of Biogeography*, 47, 181–192.

Peixoto, T.P. (2014). Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X*, 4, 011047.

Peixoto, T.P. (2017). Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95, 012317.

Peixoto, T.P. (2021). Descriptive vs. Inferential community detection in networks: Pitfalls, myths, and half-truths. *arXiv.org*.

Yen, T.-C. & Larremore, D.B. (2020). Community detection in bipartite networks with stochastic block models. *Physical Review E*, 102, 032309.