FINITE-STATE PARAMETER SPACE MAPS FOR PRUNING PARTITIONS IN MODULARITY-BASED COMMUNITY DETECTION

Ryan A. Gibson, Peter J. Mucha

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

We develop and demonstrate a method for pruning sets of network partitions to identify small subsets that are significant from the perspective of stochastic block model inference. Crucially, our method works for single-layer and multi-layer networks, as well as for restricting focus to a fixed number of communities when desired. We additionally implemented a Python package, which is available at https://github.com/ragibson/ModularityPruning.

Equivalence Between Modularity Maximization and Maximum Likelihood Methods

One of the most popular methods for community detection is to heuristically maximize a quantity known as modularity, which is usually defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \gamma \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \,, \tag{1}$$

where A is the adjacency matrix of the network, m is the number of edges, k_i is the degree of node i, and c_i is the community label of node i. Here, γ is a resolution parameter introduced by Reichardt and Bornholdt [4] to overcome issues resolving communities in large networks.

Another popular method for detecting communities is to fit a particular generative model known as a "stochastic block model" (SBM) to the network of interest. Importantly, this method is statistically principled rather than being ad hoc or motivated through heuristics alone.

In [2], Newman showed that the maximization of (1) is identical to optimizing the likelihood fit of a degree-corrected planted partition SBM when

$$\gamma = \frac{\omega_{\text{in}} - \omega_{\text{out}}}{\ln \omega_{\text{in}} - \ln \omega_{\text{out}}} \tag{2}$$

and all communities in the model share the same in-group connection propensities ω_{in} and between-group propensities ω_{out}. We call the value in (2) the "gamma estimate" for a partition when we determine ω_{in} and ω_{out} from the empirically observed values for that partition. Notably, if a partition is optimal with respect to modularity at its gamma estimate, then it is also the best possible fit to an underlying stochastic block model.

Pruning Strategy

We use this as the basis of our pruning strategy by noting that if a partition σ_1 has a lower modularity score than σ_2 at a resolution parameter value γ, then σ_1 is a worse fit than σ_2 to all SBMs satisfying the relation in (2). In this sense, we can identify the most "important" partitions by identifying those that maximize modularity at their observed gamma estimates. We say that such a partition is "stable" under the gamma estimation map. We visualize the isolation of stable partitions in Figure 1.

This fundamentally relies on being able to efficiently identify ranges of the resolution parameter γ for which each partition has a higher modularity score than all other partitions of interest. Weir et al. [5] described the "CHAMP" (Convex Hull of Admissible Modularity Partitions) algorithm to achieve this goal. In short, CHAMP exploits the fact that modularity is linear in the resolution parameters to reduce the problem to halfspace intersection.

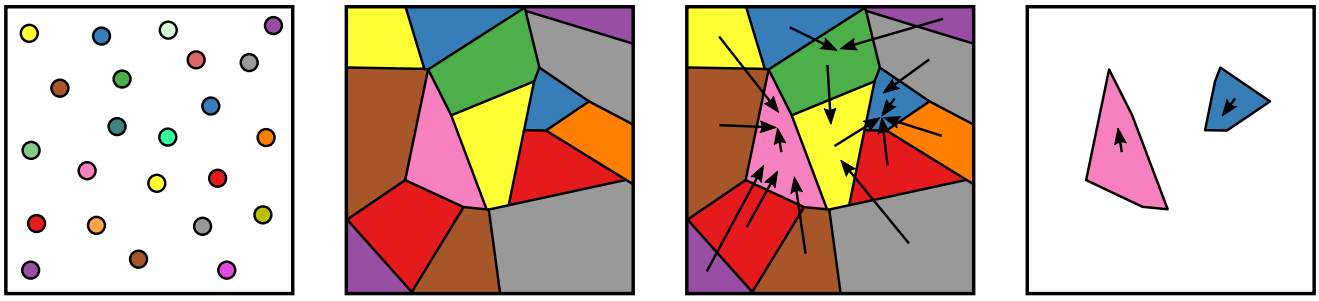
We also use (2) to derive upper bounds on the resolution parameter for which modularity maximization can be equivalent to optimizing the likelihood fit to a degree-corrected SBM, providing *a priori* regions wherein community detection heuristics "should" be run if a certain number of communities is desired.

Recently, Pamfil et al. [3] generalized Newman's equivalence to several multi-layer network models in which a collection of interrelated networks are treated as individual "layers" in a larger, connected network. We use this to extend our pruning strategy to multi-layer networks.

Results

We have demonstrated our method on the Karate Club for ease of explanation and have tested on several multi-layer networks, finding that we are often able to prune the number of identified partitions by orders of magnitude.

For instance, we found more than 500 unique partitions by running the Louvain algorithm [1] across a range of



(a) Input partitions obtained at different parameter values (b) CHAMP: removing the nowhere dominant partitions (c) Parameter estimation map on CHAMP domains (d) Pruning to the "stable" partitions (fixed points)

Figure 1: Visualization of our method. (a) Input partitions are obtained, usually through modularity maximization procedures at various points across the resolution parameter space of interest. (b) CHAMP is used to determine partitions' domains of optimality within the resolution parameter space. Partitions that are never optimal are discarded. (c) For each remaining partition, the "correct" value of the resolution parameters is estimated. Here, we depict this by drawing arrows from the partition domains to their resolution parameter estimates. (d) We return the "stable" partitions, those whose resolution parameter estimates fall within their domains of optimality; that is, we isolate the fixed points of the parameter estimation map.

values $\gamma \in [0, 2]$ on the Karate Club. Of these, CHAMP identified 9 partitions that are dominant for some value of the resolution parameter and only one of these partitions is "stable" in the sense defined in our method here. The partitions' domains of optimality and associated gamma estimates are shown in Figure 2. Here, the stable partition has four communities and a gamma estimate of $\gamma \approx 1.1$.

Similarly, in a synthetic multi-layer network, we were able to reduce a set of more than 30,000 unique identified partitions to 6 stable partitions. Notably, the stable partitions produced by our method on these synthetic networks have high alignment with the planted partition, even in "hard regime" cases where Pamfil et al.'s iterative scheme fails to converge. On a real-world multi-layer network of relationships between attorneys in a law firm, we managed to prune more than 200,000 unique identified partitions down to 3 stable partitions. Moreover, the stable partitions appear to remain high quality when the number of input partitions is significantly reduced, suggesting that our scheme can be used efficiently in practice.

By combining the ideas in [2, 3] with [5], our method addresses the problem of selecting resolution parameters (and interlayer couplings in multilayer networks) and the challenges of stochasticity due to pseudorandom computational heuristics for modularity maximization. We note that our method is agnostic to the manner in which candidate partitions are obtained and can be used to combine results from multiple heuristics.

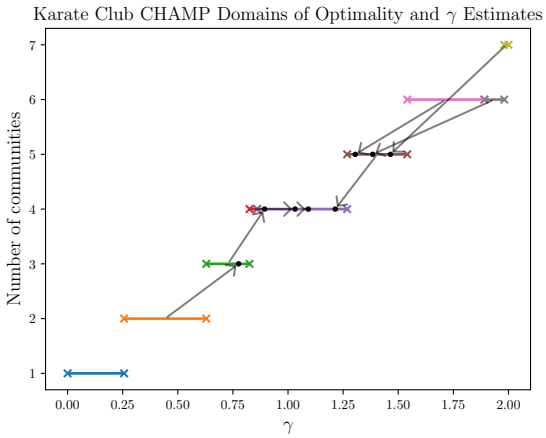


Figure 2: The domains of optimality and associated γ estimates for the partitions of the Karate Club in the pruned subset from CHAMP.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct. 2008.

[2] M. E. J. Newman. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94(5):052315, Nov. 2016.

[3] A. R. Pamfil, S. D. Howison, R. Lambiotte, and M. A. Porter. Relating modularity maximization and stochastic block models in multilayer networks. Apr. 2018.

[4] J. Reichardt and S. Bornholdt. Statistical Mechanics of Community Detection. Mar. 2006.

[5] W. H. Weir, S. Emmons, R. Gibson, D. Taylor, and P. J. Mucha. Post-Processing Partitions to Identify Domains of Modularity Optimization. *Algorithms*, 10(3):93, Sept. 2017.