Finite-State Parameter Space Maps for Pruning Partitions in Modularity-Based Community Detection

Ryan A. Gibson[1,2,*] and Peter J. Mucha[1,3]

[1]Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599-3250, USA

[2]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599-3175, USA

[3]Department of Applied Physical Sciences, University of North Carolina, Chapel Hill, NC 27599-3050, USA

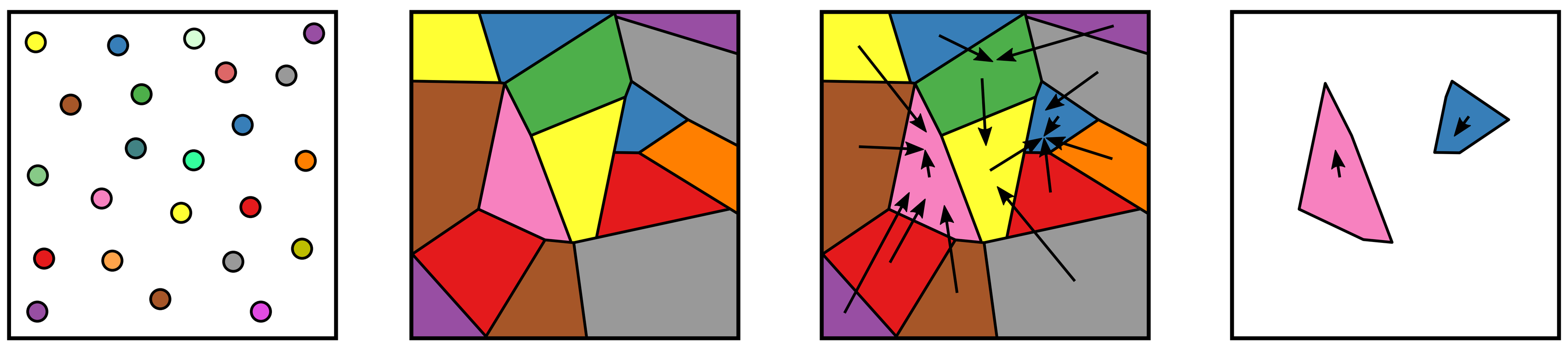
[*][EMAIL OMITTED]

1. Introduction

- Despite some known limitations, modularity maximization remains one of the most popular methods of community detection.
- In [1], Newman shows that maximizing modularity $Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \gamma\frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$ becomes equivalent to the maximum likelihood fit to a planted partition SBM when $\gamma = \frac{\omega_{\text{in}} - \omega_{\text{out}}}{\ln\omega_{\text{in}} - \ln\omega_{\text{out}}}$. When ω_{in} and ω_{out} are empirically estimated, we call this value the "gamma estimate" of a partition.
- By combining Newman's equivalence, Pamfil et al.'s [2] extension of the equivalence to several multi-layer network models, and the CHAMP partition post-processing algorithm of Weir et al. [3], we develop a method for pruning sets of network partitions to identify small subsets that are significant from the perspective of stochastic block model inference.
- This provides a procedure for exploring the resolution parameter space in modularity-based community detection and addresses the challenges of stochasticity due to pseudorandom computational heuristics. We additionally implemented a Python package, which is available at `https://github.com/ragibson/ModularityPruning`.

2. Our Method

- Note 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 gamma estimate relation.
- We propose a pruning scheme that identifies the most "important" partitions by isolating those that maximize modularity at their observed gamma estimates.

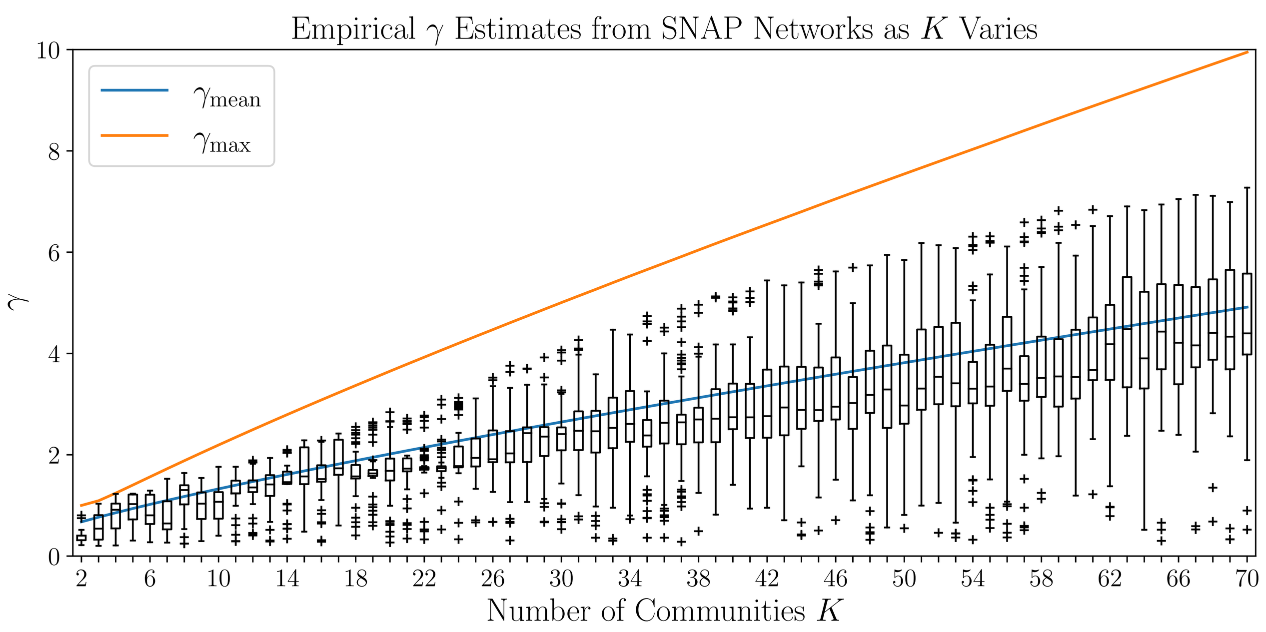


(a) Input partitions obtained at different parameter values **(b)** CHAMP: removing the nowhere dominant partitions **(c)** Parameter estimation map on CHAMP domains **(d)** Pruning to the "stable" partitions (fixed points)

4. Maximum γ estimates

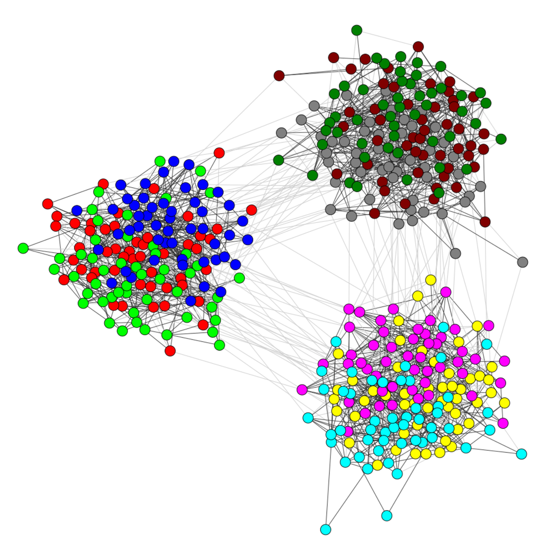
- We use the gamma estimate formula to derive upper bounds on the resolution parameter for which modularity maximization can be equivalent to assortative, degree-corrected SBM inference.
- This provides *a priori* regions wherein community detection heuristics "should" be run if a certain number of communities is desired. For example, above $\gamma \approx 1.0926$, modularity maximization is only equivalent to planted-partition, degree-corrected stochastic block model inference of four or more blocks.
- Below: Some of these $\gamma_{\max}$ bounds and the observed γ estimates on 16 social networks from the Stanford Large Network Dataset [4]. Here, box plots collect γ estimates from 1000 Louvain runs on a uniform $\gamma \in [0, 10]$ grid on each network, grouped by K.

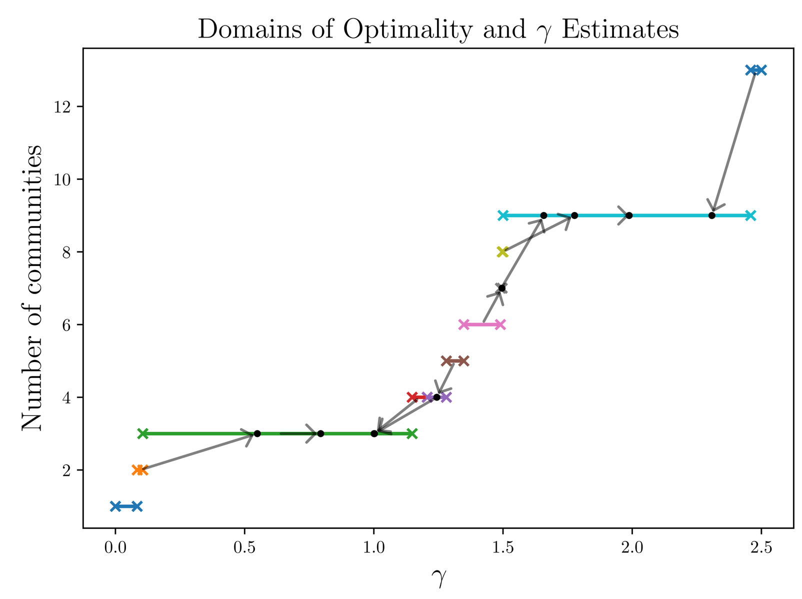
K	$\gamma_{\max}$
2	1.0000
3	1.0926
4	1.2427
5	1.4027
6	1.5640
7	1.7241
8	1.8824
9	2.0388
10	2.1931



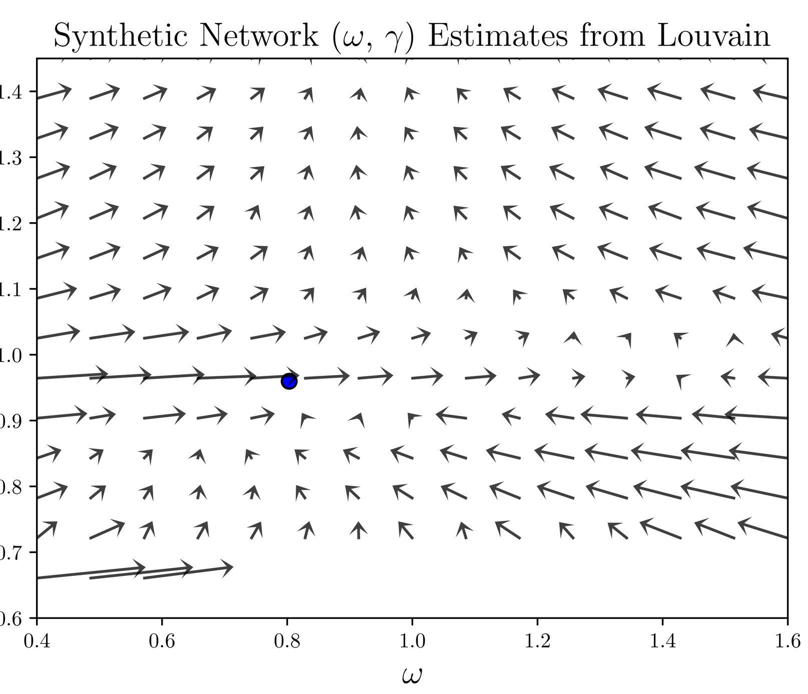
3. Selected Results

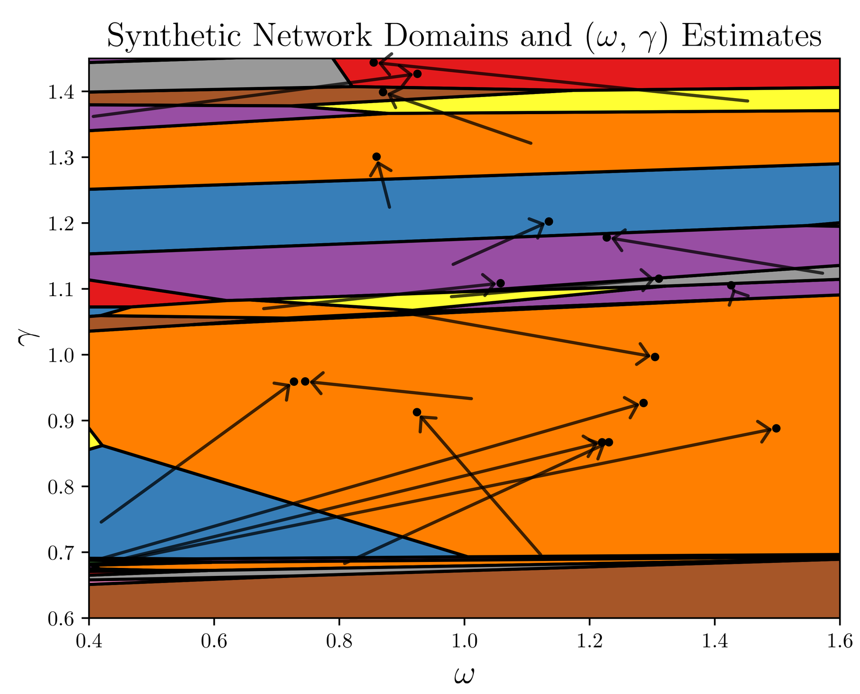
- On an SBM with nested community structure (3 large communities, each composed of 3 smaller communities), our pruning method recovers stable partitions at both the 3-community and 9-community scales.
- Left: force-directed layout of one realization of the SBM with the 9 block planted community structure colored.
- Right: results from our pruning method on 10,000 Louvain runs across $\gamma \in [0, 2.5]$. We find two stable partitions (one with 3 communities, one with 9 communities) which are very highly aligned with the planted SBM community structure.





- On a realization of a "hard regime" synthetic multi-layer network as described in Pamfil et al. [2], our method shows improvement over their proposed iterative method.
- Left: rough behavior of Pamfil et al.'s iterative estimation scheme. As in [2], it fails to converge to the ground truth community structure at $(\omega, \gamma) \approx (0.80, 0.96)$.
- Right: When input with the results of Louvain runs on a 255×255 uniform grid of $\omega, \gamma \in [0, 2]$, our method prunes the ~50K partitions to 3 stable partitions. One of these has a very large domain of optimality and high alignment with the ground truth community structure. We find similarly high quality results when reducing the number of Louvain runs to ~50, suggesting that our method can be efficient in practice.





5. References

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

[2] A. Roxana Pamfil, Sam D. Howison, Renaud Lambiotte, and Mason A. Porter. Relating Modularity Maximization and Stochastic Block Models in Multilayer Networks. *SIAM Journal on Mathematics of Data Science*, 1(4):667–698, January 2019. Publisher: Society for Industrial and Applied Mathematics.

[3] William H. Weir, Scott Emmons, Ryan Gibson, Dane Taylor, and Peter J. Mucha. Post-Processing Partitions to Identify Domains of Modularity Optimization. *Algorithms*, 10(3):93, September 2017.

[4] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. June 2014.