Adaptive Multiscale for Resolving Modularity Anomalies

B. Cross, X. Lu, B. Szymanski

Rensselaer Polytechnic Institute

October 19, 2022

The Problem

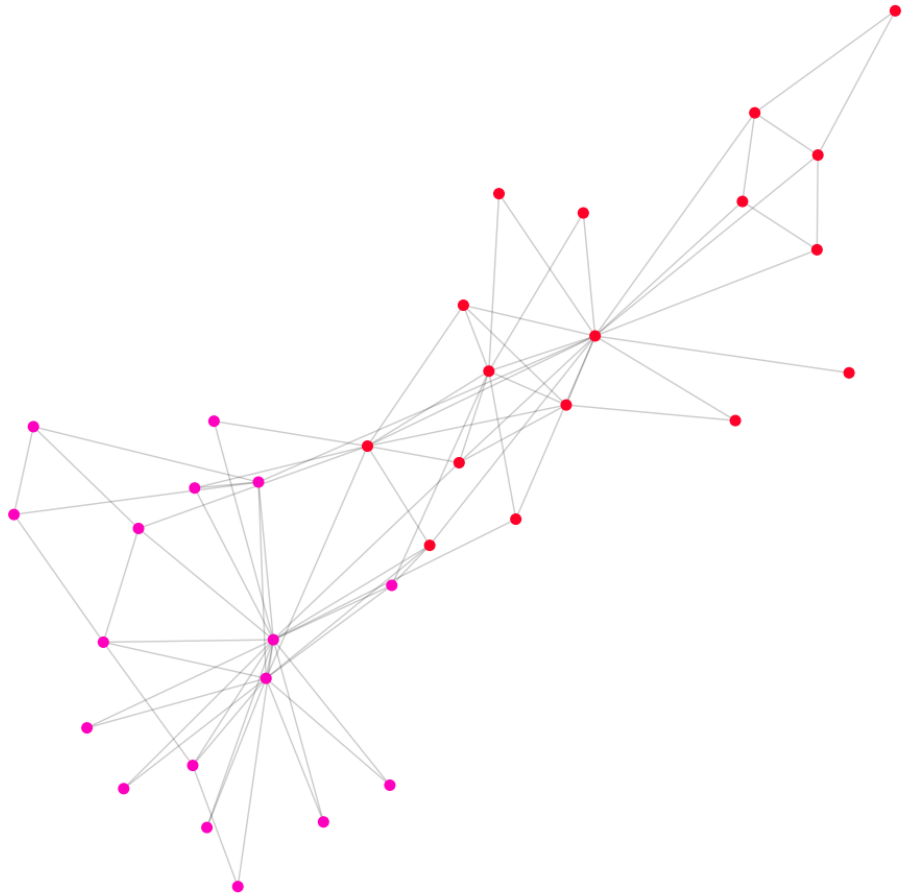
Community Detection in Heterogeneous Networks:

- We want to detect community structures accurately in networks with heterogeneous community structures.
- To explore this, we will discuss what makes a network heterogeneous in structure, why this is an issue for traditional community detection algorithms, and how we solve for it.

Background

Community Detection:

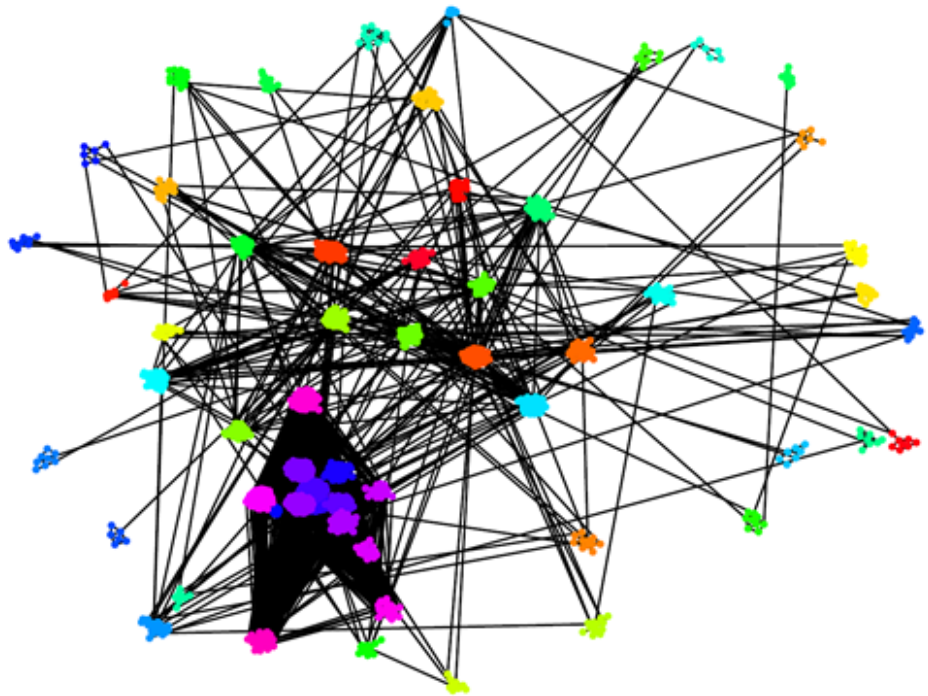
- Complex networks may be partitioned into sets of nodes that are better connected to each other than they are to the rest of the network.
- These communities provide insight into the structure of the network and the interaction between its components:
  - Communities of friends in a social network
  - Groups of proteins that interact with each other
  - etc...



Background

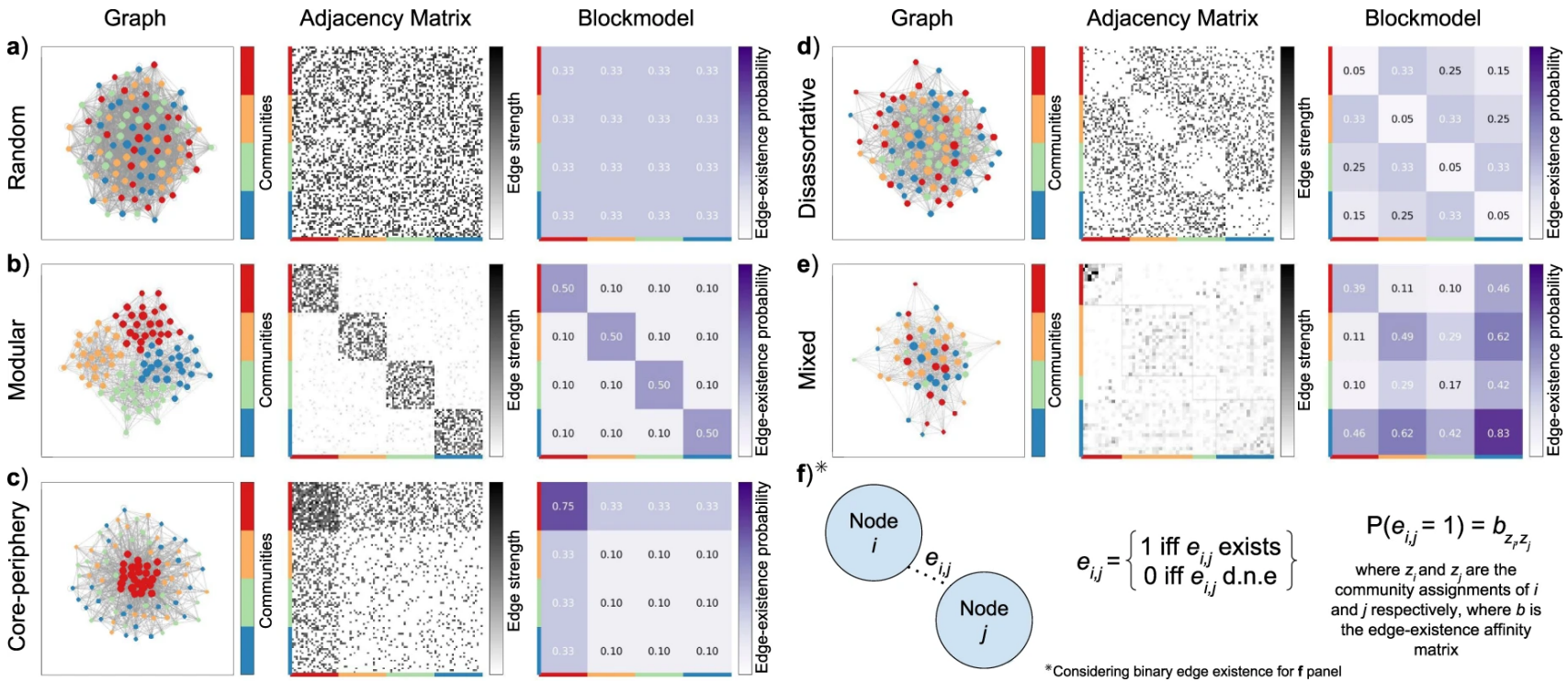
Heterogeneous Communities

- Characterized by communities that vary in size, in/out degree, connectivity characteristics, etc.
- Most community detection methods tend to detect communities that are statistically similar, this can lead to improperly merging or splitting communities.



Background

Stochastic Block Model: A generative model for networks, we define a block matrix that defines a set of connection probabilities for nodes within each of the blocks towards any other block.



[1]

[1] Faskowitz, Joshua, et al. "Weighted stochastic block models of the human connectome across the life span." Scientific reports 8.1 (2018): 1-16.

Background

Modularity: One popular method for community detection is the maximization of a metric known as Modularity.

- Modularity, defined as:

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \tag{1}$$

- This is the sum, over communities in the network, of the edge density within this community vs the expected edge density for those nodes randomly reconfigured.
- A higher modularity indicates a better partition of the network.

Modularity maximization is a fast and very effective way to partition networks, and therefore is used often in the context of large networks.

Generalized Modularity

- A more general form of Modularity exists, defined by Reichardt and Bornholdt, that takes in an additional parameter γ:

$$Q(\gamma) = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \gamma \left(\frac{d_s}{2L} \right)^2 \right]$$

- This additional parameter control's the "resolution" of the communities, where at low values we get larger communities and at high values we get smaller.

Modularity Anomalies

Resolution Limit: Modularity maximization is not without flaw, and the most well known is the resolution limit.
Using (1) we can derive a formula for the change in modularity from merging communities r and s as so:

$$\Delta Q_{rs} = \frac{m_{rs}}{m} - \frac{k_r k_s}{2m^2} \tag{2}$$

where m is the total edges in our network, m_{rs} is the number of edges between communities r and s, and k_r is the degree of community r.
Here we see that if the RHS becomes less than the LHS, then the change in modularity will be positive for any number of links between r and s greater than 0. If we consider very large networks, like WWW or social networks, this is a big issue.

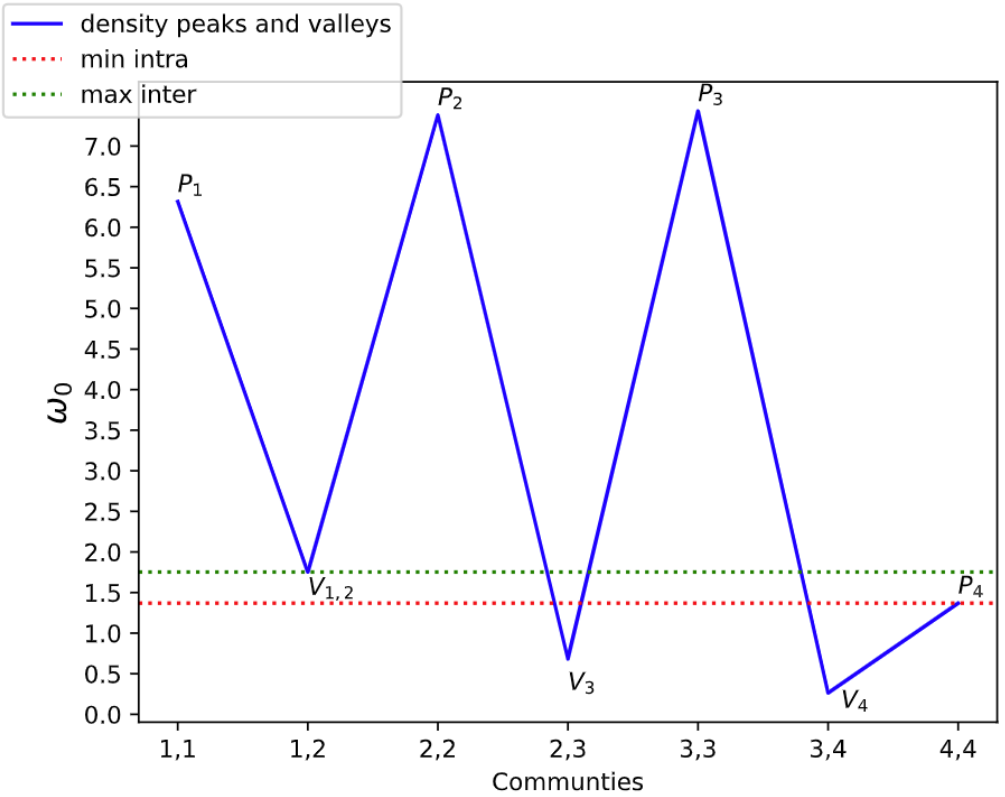
Modularity Anomalies

Extreme Example:

- The ring of cliques is the classic example, each clique is an obvious example of a community, but if we have many cliques in a ring, the maximum modularity partition is one that merges adjacent cliques.

Plateau Problem

The plateau problem is an extension of the resolution limit issue, and for it we consider the community density matrix:

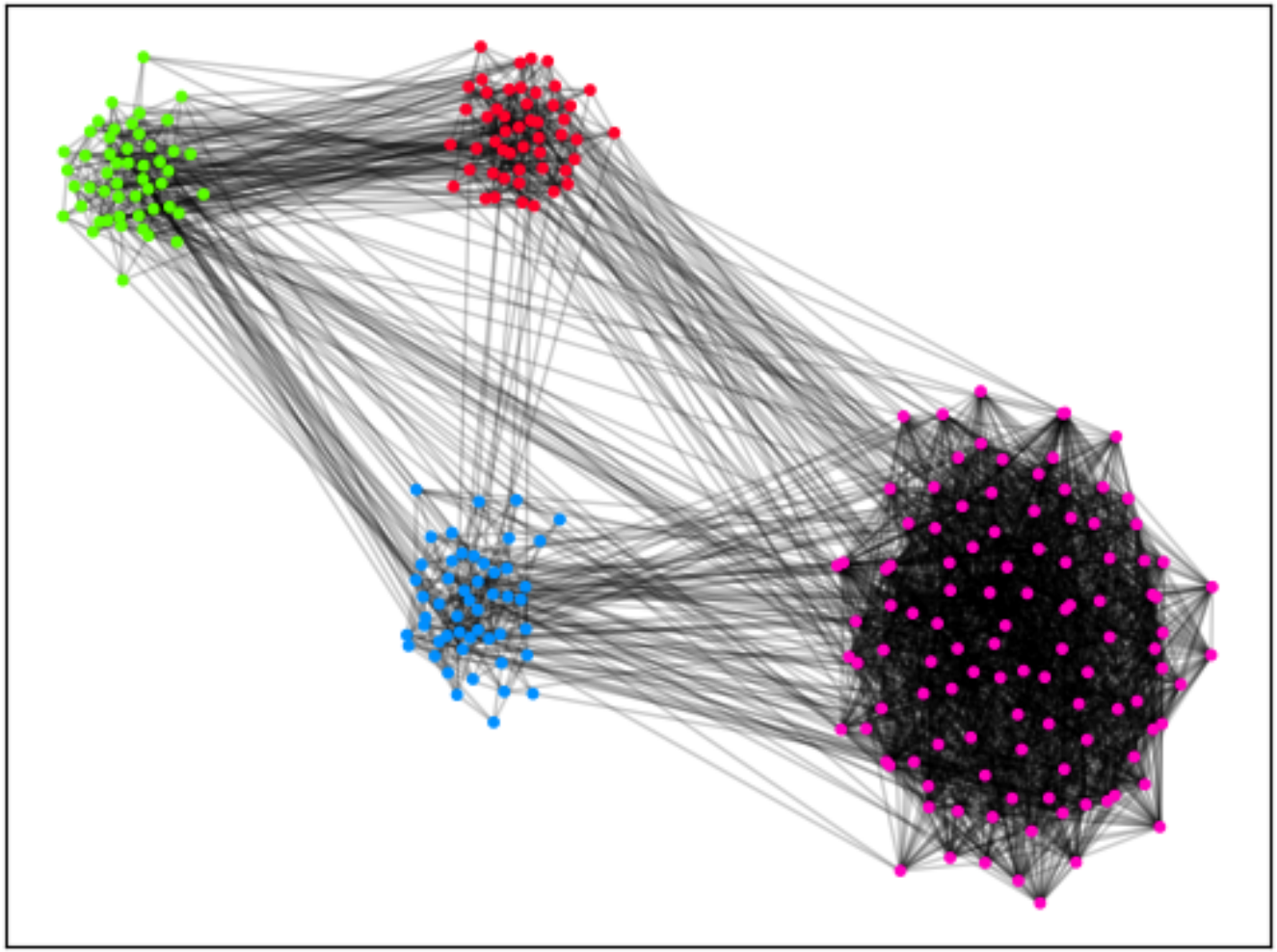


(a) Peaks and valleys

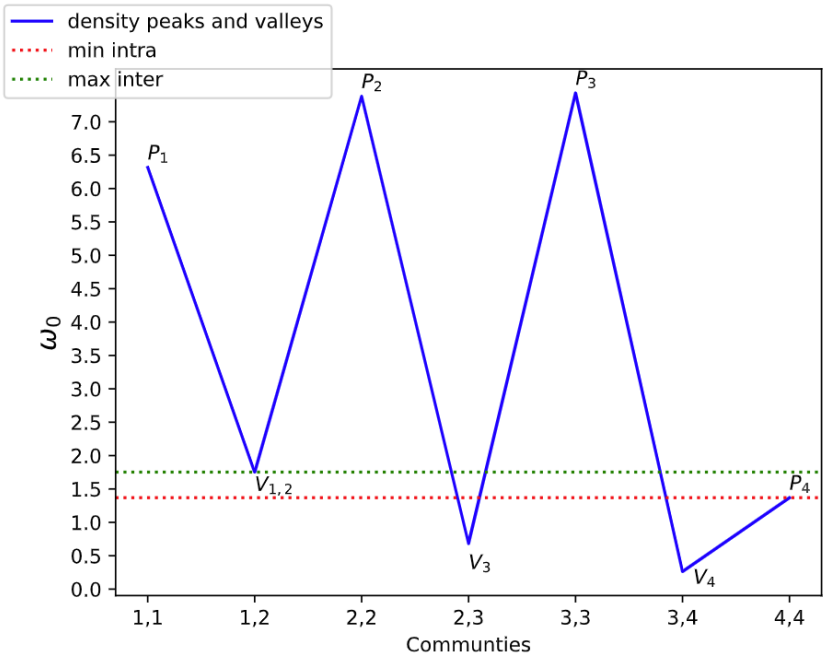
$$\Omega = \begin{bmatrix} 6.31 & 1.75 & 0.15 & 0.13 \\ 1.75 & 7.38 & 0.68 & 0.08 \\ 0.15 & 0.68 & 7.43 & 0.26 \\ 0.13 & 0.08 & 0.26 & 1.36 \end{bmatrix}$$

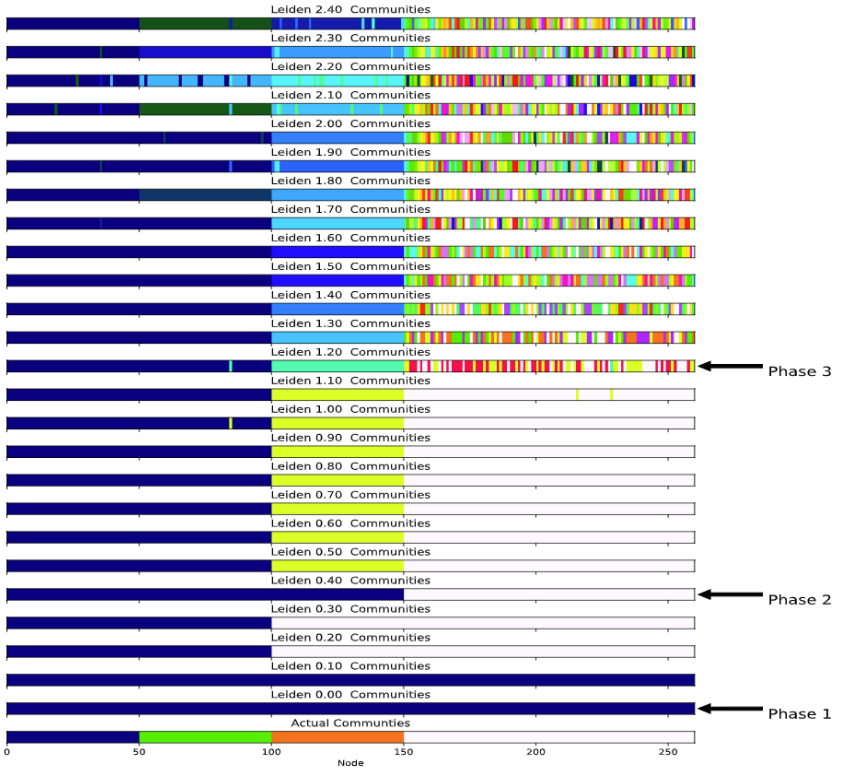
(b) Community density matrix

Plateau Problem



Plateau Problem





Leiden detected communities when resolution parameter increases from 0.0. to 2.4

No single γ can resolve all communities.

Solving the Plateau Problem

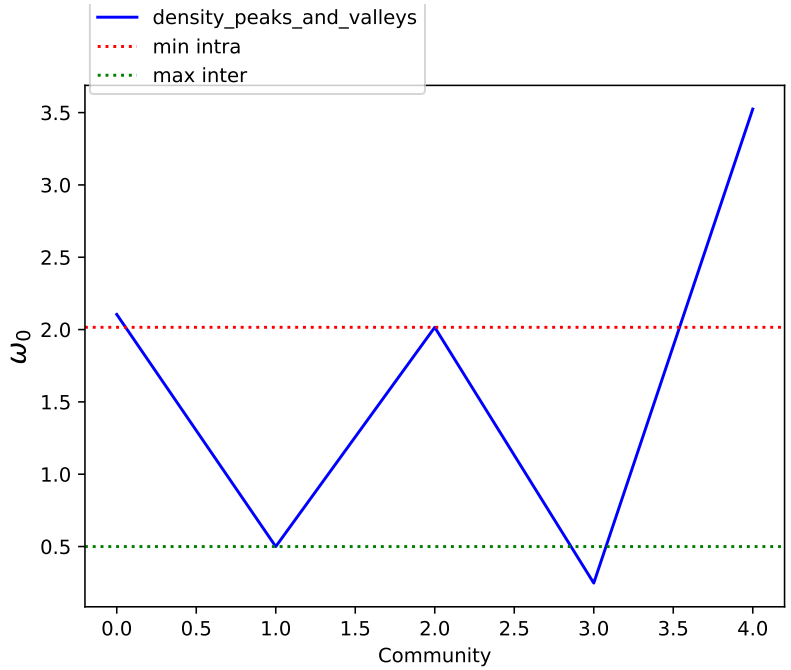
Multiscale Community Detection

- To handle the Plateau problem (Lu 2020) developed the Multiscale Community Detection Algorithm
- This heuristic algorithm applies the Louvain algorithm recursively to the initial graph and discovered community subgraphs until no good partitions of those subgraphs is found.
- The recursion terminates by Bayes model selection, which attempts to determine if the derived partition is significant by seeing if our subgraph was more likely to be generated by the configuration model than the stochastic block model.

Multiscale Algorithm

Considering our previous example, at our first level of recursion.

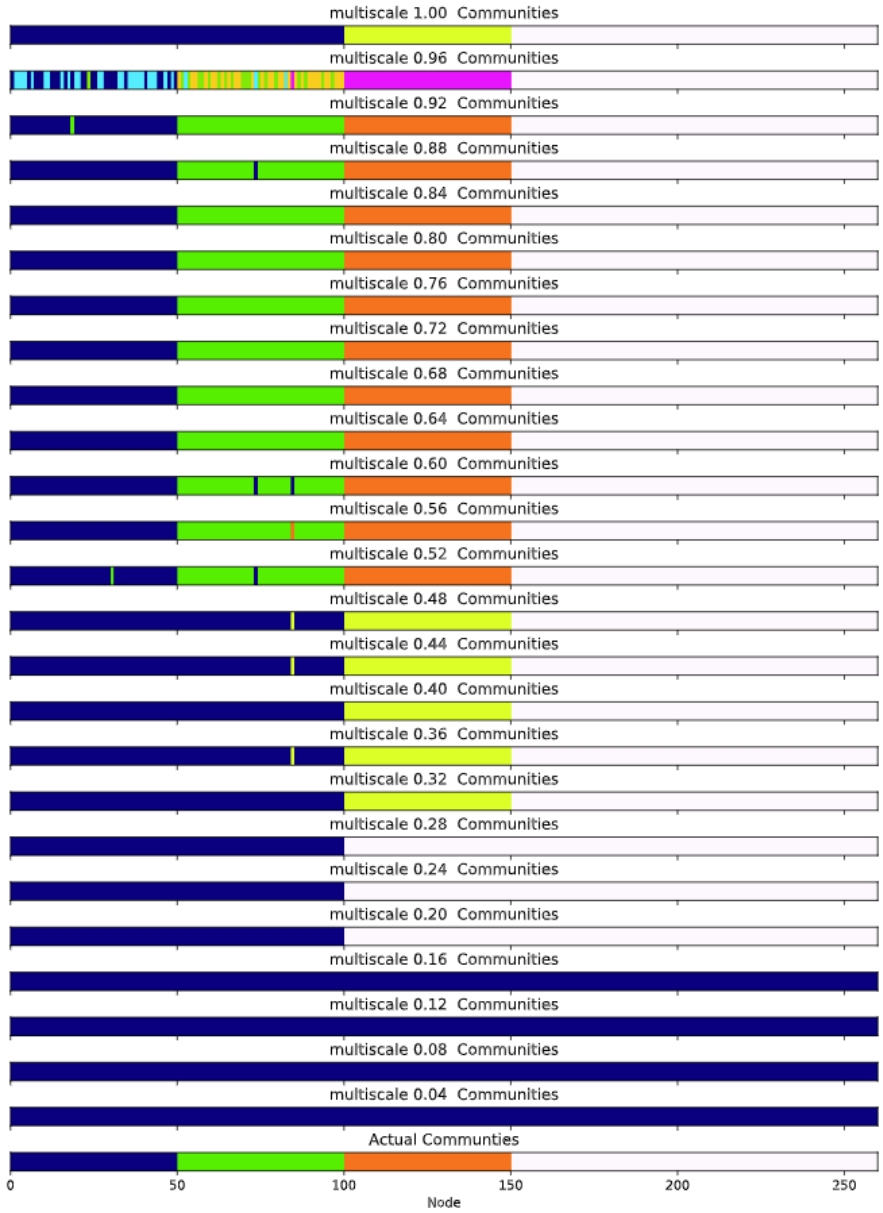
$$\Omega = \begin{bmatrix} 2.104 & 0.499 & 0.071 \\ 0.499 & 2.015 & 0.247 \\ 0.071 & 0.247 & 3.524 \end{bmatrix}$$



Our group of three communities has a density matrix that is manageable, no longer displaying the plateau problem, this is solvable by Louvain.

Multiscale Algorithm

- Multiscale is able to handle the modularity anomaly for a large range of resolutions.



Adaptive Multiscale

- The Multiscale algorithm performs well, allowing us to avoid the plateau problem, but can we improve on the algorithm?
- Potential issues:
  1) The algorithm implicitly increases the resolution parameter at each level of recursion, linearly with removed edges.
  2) The resolution parameter can be unintuitive to set properly for any arbitrary network.
  3) The Bayesian odds used to evaluate potential partitions can inadvertently allow us to improperly merge partitions.
- We can address these issues with some modifications to the Multiscale algorithm:

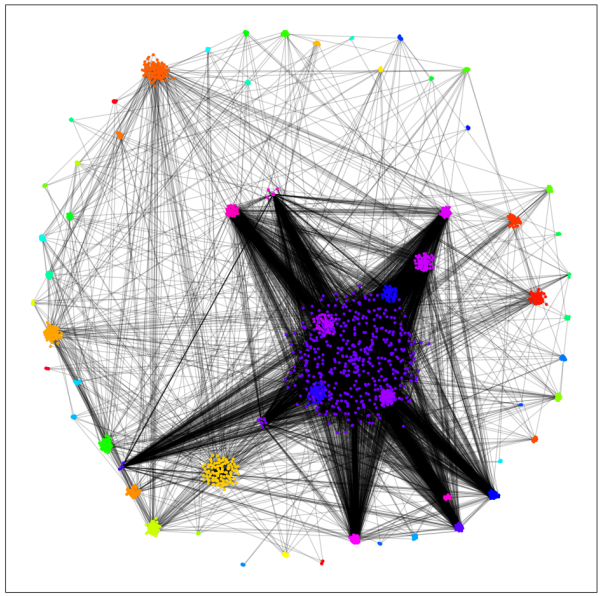
Adaptive Multiscale

Adaptive differs from multiscale in 3 important ways

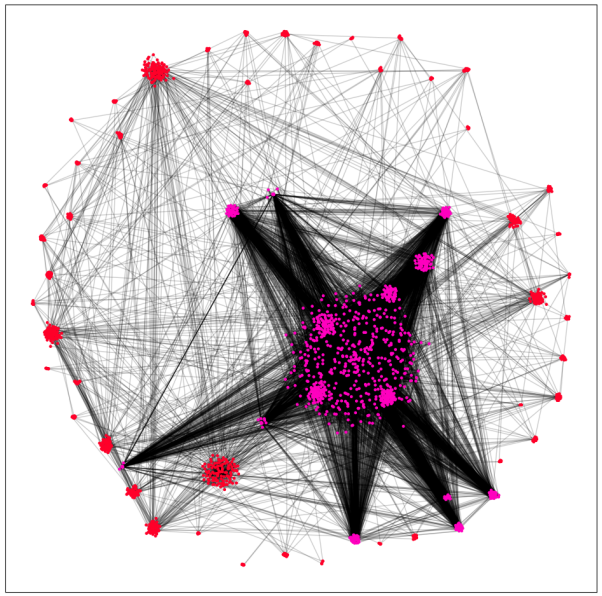
1) Dynamically selects the resolution parameter at each level of recursion.
2) Changes the termination condition of recursion, instead of stopping when a poor partition is found, keep splitting until we hit the min community size and then compare split and merge partitions as we return, selecting the best.
3) Determines the better of competing partitions.

Adaptive Multiscale

The Model selection issue: The Bayes Model Selection we use is not good at comparing the split and merge partitions of the network.



(a) Ground Truth, odds: 2.34, modularity: 0.37



(b) Merged, odds: 4.09, modularity: 0.26

Figure: Comparing partition odds

Model Selection

Our Model selection step consists of comparing the posterior probabilities of having generated our network using the planted partition model vs the null model, in this case an SBM with only one module and therefore one density parameter dictating connections between nodes.

$$\mathbf{\Lambda} = \frac{P(\mathbf{g},\mathcal{H}_1|\mathbf{A})}{P(\mathcal{H}_0|\mathbf{A})} = \frac{P(\mathbf{A}|\mathbf{k},\mathbf{g},\mathcal{H}_1)}{P(\mathbf{A}|\mathbf{k},\mathcal{H}_0)} \times \frac{P(\mathbf{g})P(\mathbf{k})P(\mathcal{H}_1)}{P(\mathbf{k})P(\mathcal{H}_0)}$$

Figure: Posterior odds

$$P(\mathbf{A}|\mathbf{k},\mathbf{g},\mathcal{H}_1) = 2\sum_r m_r \times \ln \frac{2\sum_r m_r}{\sum_r \frac{k_r^2}{2m}} + (2m - 2\sum_r m_r) \times \ln \frac{2m - 2\sum_r m_r}{2m - \frac{k_r^2}{2m}}$$

Figure: Posterior Probability Planted Partition Model

Model Selection

Work Around:

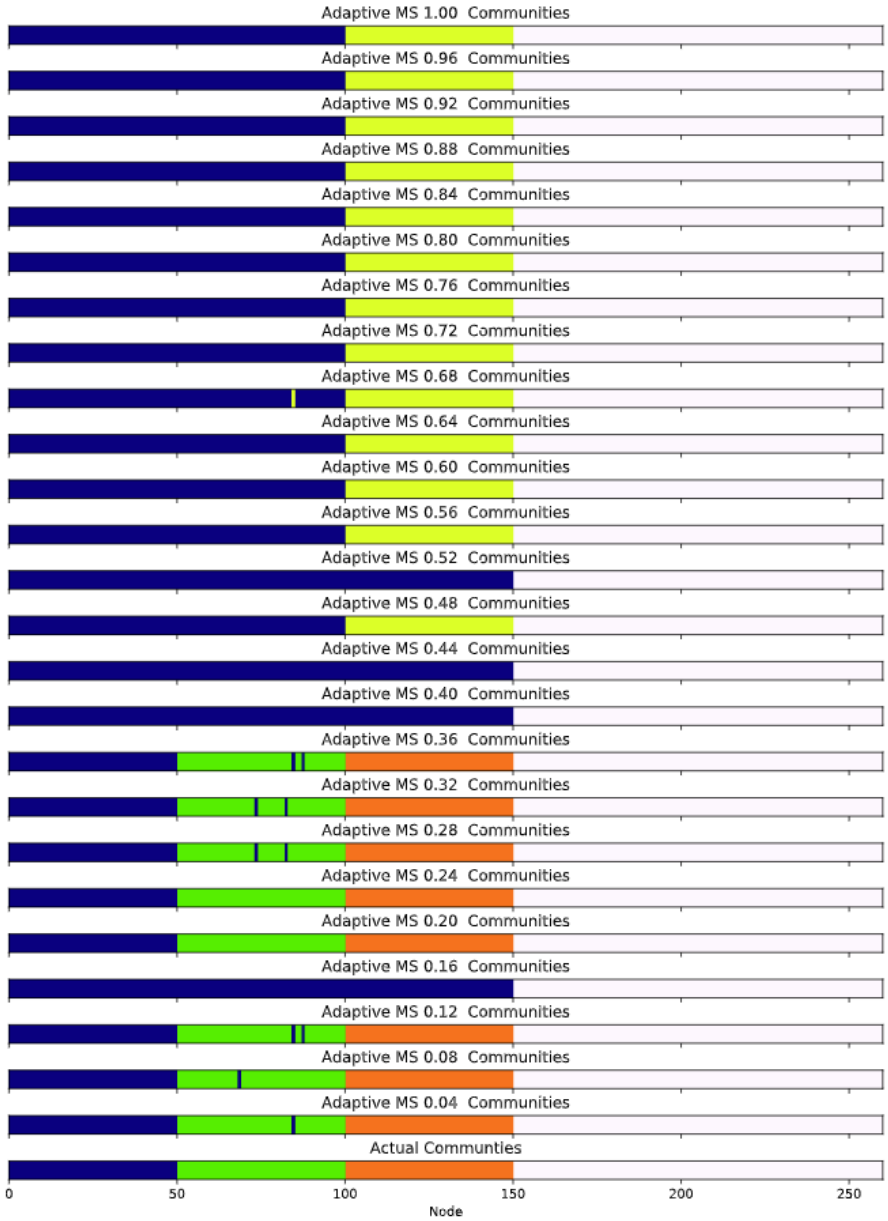
- The current workaround for the issue of poor split / merge partition comparisons is to use a tiebreaker metric when the odds of split / merge are close.
- The tiebreaker metric can be anything, modularity showed the best overall performance.
- This method is flimsy, given that modularity is directly tied to the issue we are trying to solve, but has shown good results in practice.

Testing

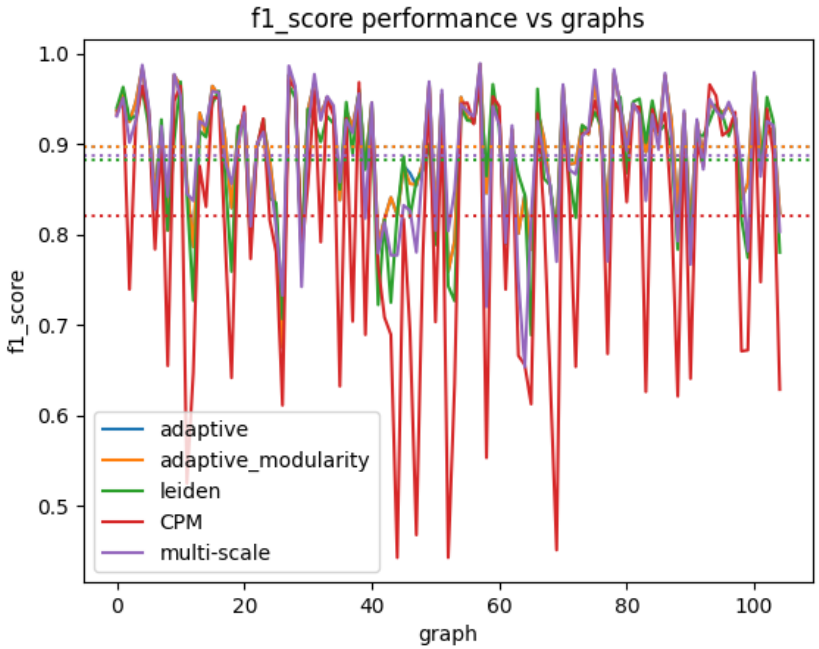
- We test the performance of our algorithm using real world and generated test networks.
- Our synthetic networks are generated to display heterogeneous community properties by re-configuring a set of base LFRs, combining them in a way that retains the internal and external density of each community, along with the degree distribution of each node.

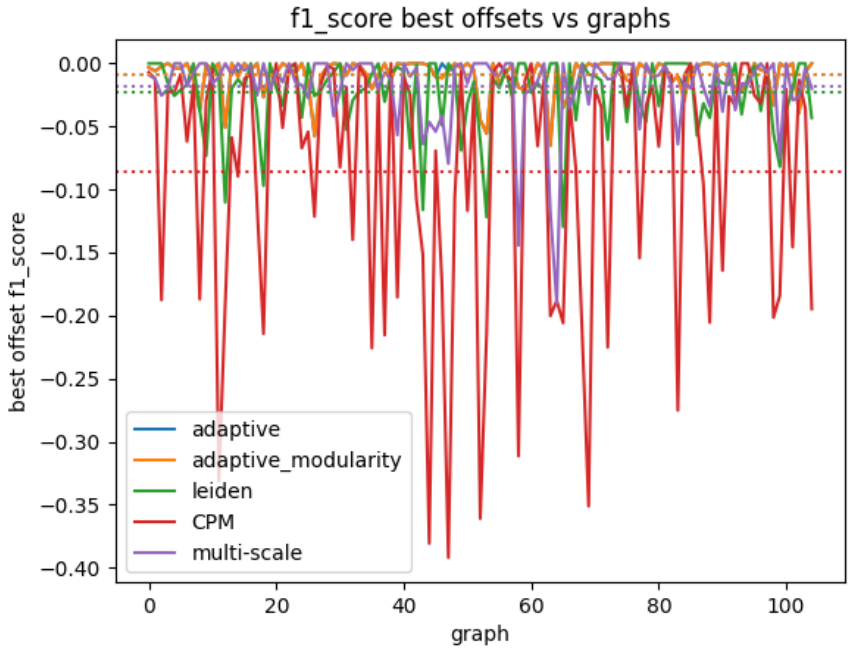
Plateau Problem Example

Adaptive is able to handle the plateau problem handily.



Generated Network Performance





Real Network Performance

Network Tested	Metric	Adaptive MS	Multi-scale	Adaptive MS-Q	CPM	Leiden
Mouse Cell	ARI	0.415	0.479	**0.507**	0.299	0.502
N=1,885	Topo NMI	0.727	0.736	**0.763**	0.676	0.758
E=40,096	F1-Score	0.624	0.673	**0.719**	0.643	0.715
Amazon	ARI	0.877	0.770	**0.942**	0.870	0.883
N=16,716	Topo NMI	0.992	0.983	**0.993**	0.991	0.992
E=48,739	F1-Score	0.968	0.908	**0.981**	0.976	0.973
DBLP	ARI	0.062	0.106	0.117	0.012	**0.168**
N=93,432	Topo NMI	0.759	0.754	**0.761**	0.749	0.691
E=335,520	F1-Score	0.436	0.426	**0.458**	0.416	0.358
Live Journal	ARI	0.717	**0.746**	0.704	0.576	0.675
N=84,438	Topo NMI	0.970	**0.973**	0.969	0.963	0.954
E=1,521,988	F1-Score	0.886	**0.915**	0.879	0.832	0.807

Table: Real Network Results

Improvement Goals

While the adaptive algorithm performs well on our synthetic and real world tests, there are some things that should be improved on to increase the stability of the algorithm over different test cases:

- Improve our model selection method so that we don't have to make use of the tie breaker metric.
- Resolution parameter selection, we can improve algorithm speed and improve the quality of our recursive partitions by intelligently choosing the resolution parameter.

Current Attempts

Nested DC-SBM: This model, proposed by Tiago Peixoto, alters the prior for the edge counts by considering a partition heirarchy.

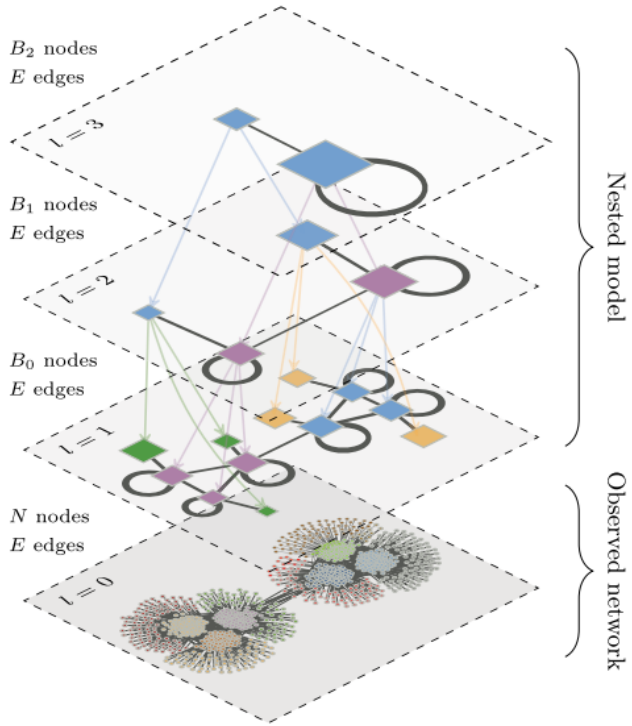


Figure: Nested DC-SBM

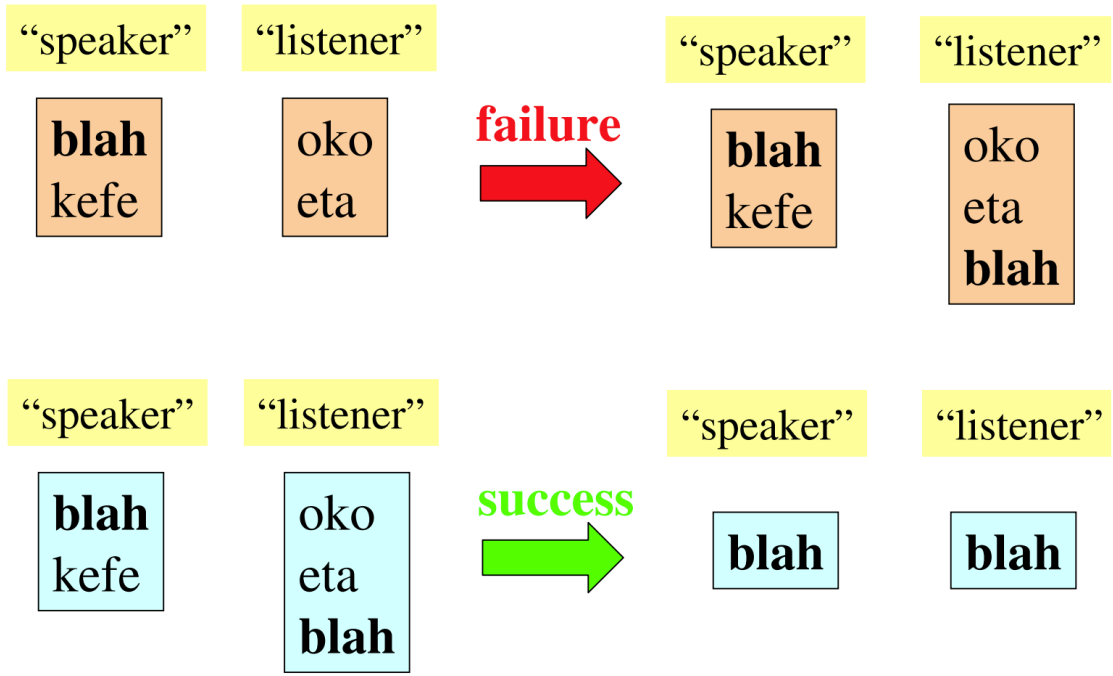
- Each level considers each group of the partition below it to be its own nodes, keeping the same number of edges as the previous level.
- This prior now includes all the information missing from our edge count prior, which only considers the densities ω_0 and ω_1.

2

[2] Peixoto, Tiago P. "Hierarchical block structures and high-resolution model selection in large networks." Physical Review X 4.1 (2014): 011047.

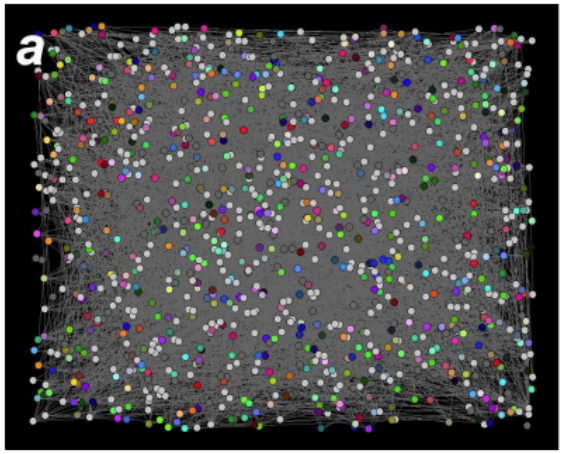
Naming Game

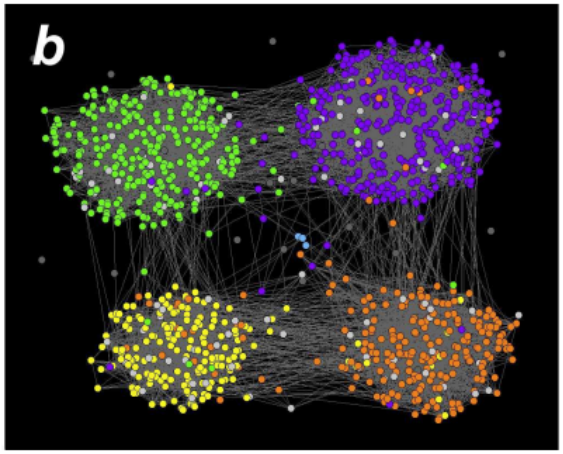
- The Naming Game is a consensus game played on a network where the goal is for each node in the network to agree on a common label.
- There has been a good deal of work done showing that community structure has strong influence on the outcome of this game

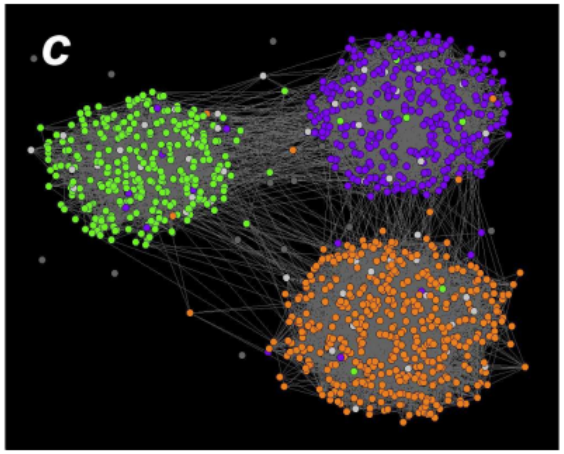


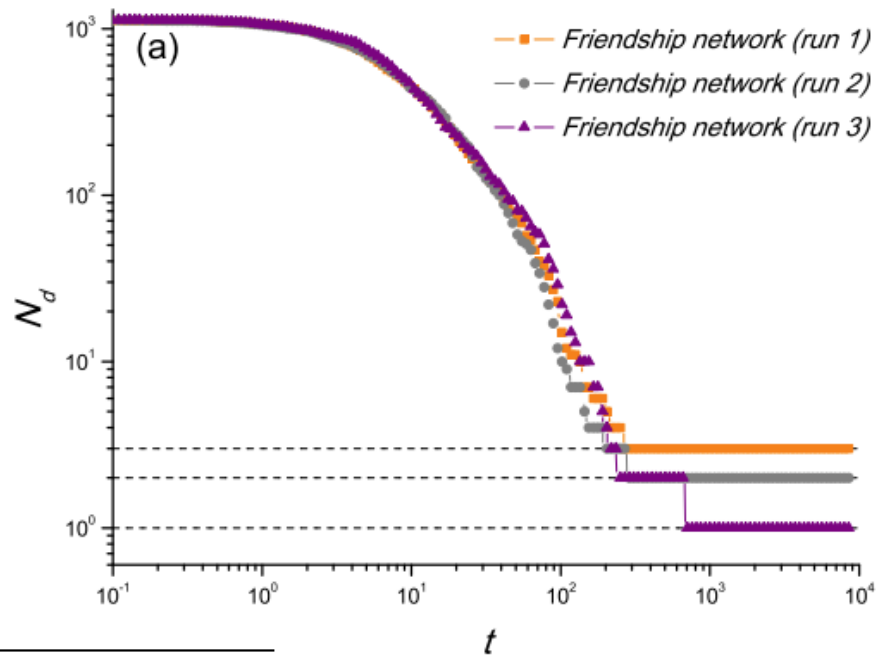
[3]Lu, Qiming, Gyorgy Korniss, and Boleslaw K. Szymanski. "The naming game in social networks: community formation and consensus engineering." Journal of Economic Interaction and Coordination 4.2 (2009): 221-235.

Naming Game









[4]

[4] Lu, Qiming, Gyorgy Korniss, and Boleslaw K. Szymanski. "The naming game in

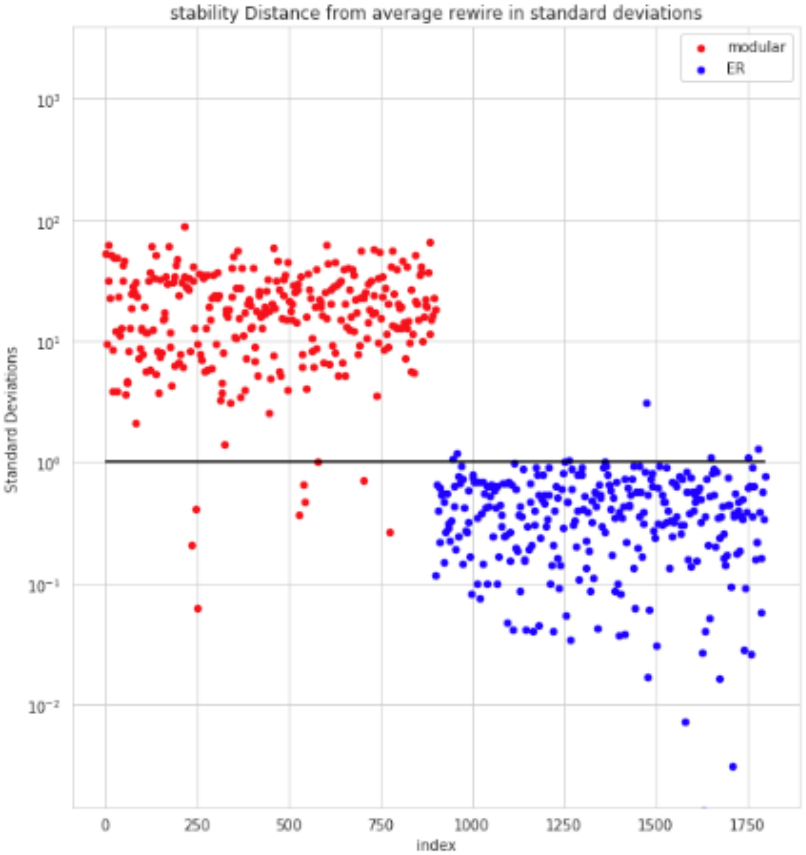
Naming Game

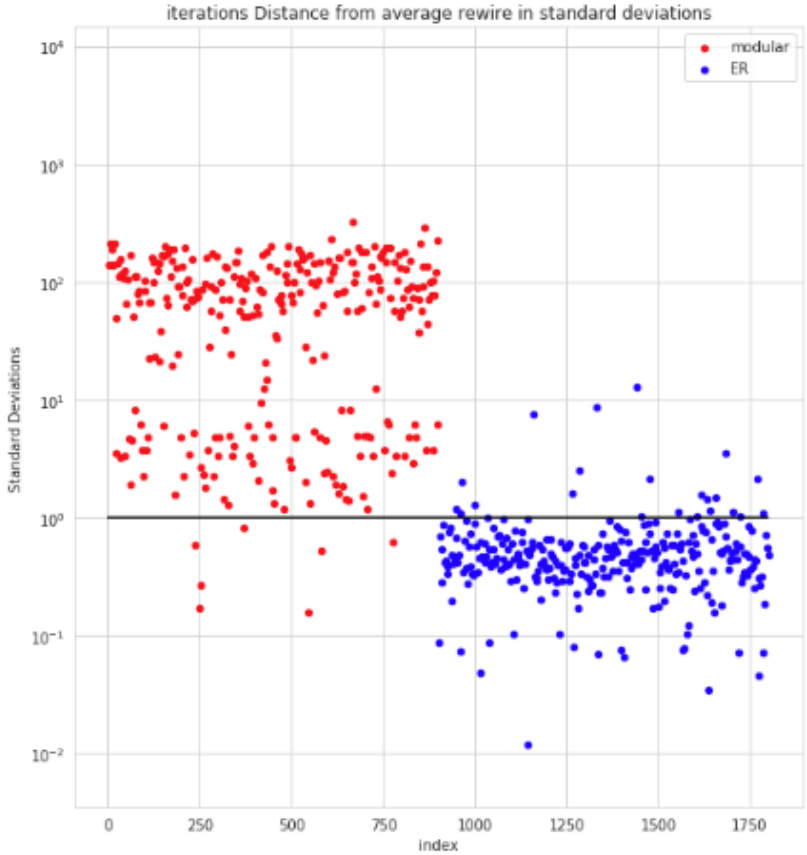
Utilizing the naming game for community detection

- Average naming game performance on target network
- Randomly rewire the network (Configuration Model), get average NG performance
- Compare the performance of the two methods, if different enough then we expect community structure

Naming Game

Performance on networks with and without communities:





Conclusion

- Adaptive modularity shows strong performance over a wide array of homogeneous and heterogeneous networks.
- Still work to be done with model selection improvements
- A sound improvement to the model selection step would greatly increase the stability of the algorithm.

Citations

- Lu, X., Cross, B., Szymanski, B. K. (2020). Asymptotic resolution bounds of generalized modularity and multi-scale community detection. Information Sciences.
- Chen, M., Kuzmin, K., Szymanski, B. K. (2014). Community detection via maximization of modularity and its variants. IEEE Transactions on Computational Social Systems, 1(1), 46-65.
- Fortunato, S., Barthelemy, M. (2007). Resolution limit in community detection. Proceedings of the national academy of sciences, 104(1), 36-41.
- Reichardt, J., Bornholdt, S. (2006). Statistical mechanics of community detection. Physical review E, 74(1), 016110.
- 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(10), P10008.

Citations

- Traag, V. A., Van Dooren, P., Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. Physical Review E, 84(1), 016114.
- Peixoto, Tiago P. "Hierarchical block structures and high-resolution model selection in large networks." Physical Review X 4.1 (2014): 011047.
- Faskowitz, Joshua, et al. "Weighted stochastic block models of the human connectome across the life span." Scientific reports 8.1 (2018): 1-16.