Effects of Null Model Choice on Modularity Maximization

Christopher Brissette, Ujwal Pandey, George M. Slota
Rensselaer Polytechnic Institute, {brissc,pandeu,slotag}@rpi.edu

CD and Modularity

Community Detection is a well-studied problem within network science.

- Essentially, community detection attempts to find relatively "dense" clusters in a network.
- To do so, one can optimizes community assignments relative to possible objective measurements.
- *Modularity* is one common objective measure.

Modularity is a measure of, given community assignment, how much more "dense" these assignments are on the given network relative to a network with randomly assigned edges.

- In our paper, we study the effect of changing the method for this relative measurement by changing the *null model* describing the "randomly assigned edges".

Contributions

Our work has several contributions towards both random graph generation and analysis as well as community detection.

1. We are the first work to extensively study the usage of a more appropriate null graph model within the context of modularity maximization.
2. We detail our approach to computing attachment probabilities and their effective utilization within a modularity maximization framework.
3. We observe that this change in attachment probabilities can improve computed NMI scores by up to fifty percent on average for some data sets.
4. We discuss how this work might be applied in future efforts, such as with multi-level community detection algorithms.

Our code and methods are available upon request!

Null Graph Model

The typical null model used for modularity is the so-called Chung-Lu model:

$$p_{u,v} = \frac{d_u d_v}{2m}$$

Here, the model describes the connection probability between nodes u, v and the product of their degrees $(d_u d_v)$ divided by two times the number of edges in the network (i.e., the degree sum over all nodes).
Using this model, modularity Q can be defined on a graph $G = (V, E)$ with a set of Communities C as the following:

$$Q = \frac{1}{2m} \sum_{u,v \in V} \left[A_{uv} - \frac{d_u d_v}{2m} \right] \delta(c_u, c_v)$$

Here δ is the kronecker delta, A is the adjancency matrix representation of G, and $fracd_u d_v 2m$ is our Chung-Lu probability of a connection between nodes $u, v \in V(G)$. In our null models we propose, we seek to replace the Chung-lu probability term simply $p_{u,v}$, yielding:

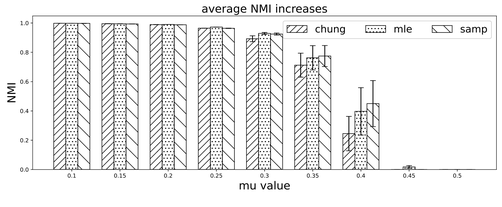
$$Q = \frac{1}{2m} \sum_{u,v \in V} \left[A_{uv} - p_{u,v} \right] \delta(c_u, c_v)$$

We can then use this generalized version of modularity with any null model of our choosing, as long as our null model can define pairwise attachment probabilities. We use two such null model choices:

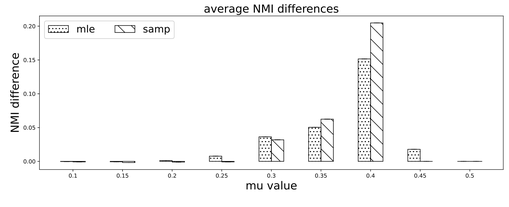
1. (*SAMP*) We define probabilities of a uniformly-random model with a fixed degree sequence. We construct the probabilities through a Markov process of double-edge swap rewiring to emperically measure average pairwise attachments over many randomly selected instances from the desired graph space.
2. (*MLE*) We note that a degree sequence can be seen as a probability distribution, and the degrees of nodes with common weights will be distributed as Poisson distributions. We can use maximum likelihood estimation to express this distribution as a sum of Poisson distributions from which nodal weights can be discerned for a Chung-Lu-like null graph model.

Results: Choice of Null Model on Community Detection NMI

We use the LFR generator, a wide suite of possible input parameters, and measure the average NMI produced through basic modularity maximization across all three test models (Chung-Lu baseline, SAMP, MLE).



The average NMI results across a general suite of LFR graphs when using our three different null models for modularity calculations.



The average relative improvement of NMI results across a general suite of LFR graphs for our two new null model choices relative to basline Chung-Lu.

Discussion of Results

Observations

- We observe NMI results for standard modularity maximization, *SAMP*, and *MLE* varying the μ parameter on LFR graphs with minimum degree 5, and minimum community size 6.
- We can see a **sharp improvement** in cluster quality near the $\mu = 0.4$ bound, implying that our new null models may perform better than standard modularity maximization in these test instances.
- *We observe similar findings across a much broader set of test instances*, as well.

Why?

- Chung-Lu probabilities can over-estimate real attachment probabilities between pairs of average degree nodes and pairs of high degree nodes within graphs with skewed degree distributions; low degree probabilities are otherwise similar.
- As a consequence, the baseline modularity maximization biases against assortativity, while most real networks and benchmark networks actually exhibit a considerable amount of assortative degree mixing within communities.
- The use of appropriate null model probabilities 're-biases' towards assortative mixing within communities when performing modularity maximization.

Future Work

Multi-level Modularity Methods

- Most modern community detection algorithms utilize a multi-level approach, where computation is accelerated by *coarsening* the graph after some number of optimization steps.
- Our null models can be applied to multi-level algorithms such as *Louvain* without much modification.