A definition of graph modularity with outliers

Bogumił Kamiński* [1], Paweł Prałat† [2], François Théberge‡ [3], and Sebastian Zając§ [1]

[1] *SGH Warsaw School of Economics, Warsaw, Poland,*
[2] *Toronto Metropolitan University, Toronto, ON, Canada,*
[3] *The Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada*

Abstract

One of the most important features of networks is their community structure, as it reveals the internal organization of nodes. Standard algorithms allowing for detection of communities usually try to assign nodes to precisely one community. On the other hand, in many complex networks, while a majority of nodes indeed can be easily associated with some communities, there is usually some subset of nodes that cannot be easily associated with a single community. As a result, there is a need to detect outlier nodes.

One of the mostly used unsupervised methods for detecting communities in graphs is maximization of the modularity function. There are several efficient approaches for community detection that use modularity, the most popular are Louvain, Leiden, and ECG.

The modularity function favours partitions of the vertex set V of a graph in which a large proportion of the edges fall entirely within the parts (often called clusters). It measures the deviation between so called edge contribution and degree tax. In our work we propose to modify the modularity function to incorporate the existence of outliers. For a given partition $\mathbf{A} = \{A_1, A_2, \ldots, A_\ell, O\}$ of V, where O is the set of outliers, the modularity function is adjusted as follows: for a given $\lambda \in \mathbf{R}_+$,

$$\begin{aligned} q_o(\mathbf{A}) \quad = \quad & \sum_{i=1}^{\ell} \frac{e(A_i)}{|E|} - \sum_{i=1}^{\ell} \left(\frac{\mathrm{vol}(A_i)}{\mathrm{vol}(V)} \right)^2 \\ & -\lambda \left[\left(\frac{e(O)}{|E|} - \left(\frac{\mathrm{vol}(O)}{\mathrm{vol}(V)} \right)^2 \right)^2 + \sum_{i=1}^{\ell} \left(\frac{e(O, A_i)}{|E|} - \frac{2\mathrm{vol}(A_i)\mathrm{vol}(O)}{\mathrm{vol}(V)^2} \right)^2 \right], \end{aligned} \tag{1}$$

where $e(A)$ is the number of edges within set A, $e(A, B)$ is the number of edges between set A and B, and $\mathrm{vol}(A)$ is the sum of degrees of nodes in A. The general idea is that nodes are identified as outliers if the distribution of their neighbours among elements of $\mathbf{A}$ is approximately equal to the expected number of neighbours in the associated random graph null model. The λ parameter allows the user to decide how strong this preference is. In particular, this model is equivalent to the standard modularity optimization when $\lambda \to +\infty$.

This work has been supported by the Polish National Agency for Academic Exchange under the Strategic Partnerships programme, grant number BPI/PST/2021/1/00069/U/00001.

*bkamins@sgh.waw.pl
†pralat@ryerson.ca
‡theberge@ieee.org
§szajac2@sgh.waw.pl