The Modularity Resolution Limit in the Malaria Antigen Gene Regulatory Network

Andy Burkhardt[1], **Max Hedeman Joosten**[1], **Lucia Pezzetti**[1], **and Lynn Müller**[1]

This manuscript was compiled on January 8, 2024

The study of network partitions into communities, which represent subgraphs where nodes exhibit stronger interconnections with each other, holds significant relevance in the realm of network theory. Modularity serves as a quality index for the partitioning of networks into communities. It measures the extent of connections within a given module concerning the expected values for a randomized graph of equivalent size and degree sequence. While modularity provides a framework for detecting communities within a network, the practical optimization of modularity does not always achieve this goal. In real-world applications, the process of modularity optimization may overlook significant substructures in a network. Consequently, it is essential to validate the modules obtained through modularity optimization. In this context, we offer an example of the malaria parasite antigen gene regulatory network to assess the reliability of this community detection method, highlighting the challenges associated with the algorithm and proposing a method of reducing the resolution limit.

Complex Networks | Modularity | Resolution Limit

Significance Statement

The analysis of the undirected biological network of recombinant antigen genes from the human malaria parasite P. falciparum provides a case study for examining the resolution limit in modularity optimization. By analyzing the community detection results in this context, we gain insights into whether modularity optimization can identify smaller structures within larger communities. The nature of antigen gene interactions, represented by the network's structure, makes it an ideal candidate for exploring how well community detection methods can capture modular patterns and whether they can reveal substructures within larger modules. This example of a real network therefore offers practical implications on why resolution limit is a problem in modularity optimization.

The paper titled "Resolution Limit in Community Detection" (1), by Fortunato & Barthélemy, delves into the limitations associated with modularity-optimization-based community detection and introduces the concept of the 'resolution limit'. In the context of community detection, the resolution limit characterizes the algorithm's capacity to identify communities within a network. It signifies the point at which modularity optimization falls short in identifying smaller structures within the network. This resolution limit is independent of the network's inherent structure and instead results from the comparison between the number of links present within interconnected communities and the total number of links within the network. In practical terms, this limitation can result in the omission of important substructures within networks.

Additionally, predicting the size of modules when using modularity optimization can be challenging, particularly for modules with only a few internal links. The text suggests that other quality functions with a similar structure, where the quality of a partition is determined by the sum of the qualities of individual modules, may face comparable resolution limits. This is due to the mathematical trade-off that underlies the origin of the resolution scale. As a result, the process of selecting the most appropriate method for community detection can be quite challenging.

The Modularity Resolution Limit

We will show the resolution limit in the same way as Fortunato and Barthélemy (1). We can write the modularity of a network partitioned into C modules as:

$$Q = \sum_{c=1}^{C} \left(\frac{L_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right) \qquad [1]$$

Where L and L_c denotes the total number of links or edges in the network and in a partition respectively. d_c equals the sum of degrees of all nodes within partition c.

Suppose that the network consists of three modules; modules M_1, M_2 and the rest of the network which we will denote by M_0. Let l_1 and l_2 denote the number of links internal to modules 1 and 2. Let $l_{int} = a_1 l_1 = a_2 l_2$ denote number of links between modules 1 and 2, with $a_1, a_2 \geq 0$. $l_1^{out} = b_1 l_1$ and $l_2^{out} = b_2 l_2$ are the number of links that enter M_0 from modules 1 and 2 respectively, with $b_1, b_2 \geq 0$.

Consider two partitions A and B, where A recognizes all modules as separate and B recognizes modules 1 and 2 as one module. Using the handshake lemma, we can express $d_c = (a_c + b_c + 2) l_c$. To properly pick the correct partition A under modularity optimization, we require that $\Delta Q = Q_B - Q_A < 0$. This is satisfied if:

$$l_2 > \frac{2La_1}{(a_1 + b_1 + 2)(a_2 + b_2 + 2)} \qquad [2]$$

It is possible to construct modules such that this inequality is not satisfied, which will be shown in the next section. This leads to a failure in detecting communities under modularity optimization.

Author affiliations: [1]Students at the Institute of Mathematics, University of Zurich

Furthermore, the above inequality depends on the size of the network through the number total number of links, L. This implies that when l_2 is small relative to L, the inequality may not hold, which means that modularity optimization fails to detect relatively small communities below a certain threshold.

A Synthethic Example

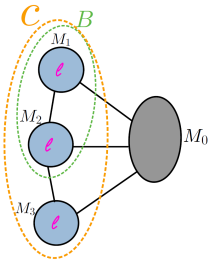

Fig. 1. Partitions

We will analyze the network shown in Fig. 1, consisting of three modules M_1, M_2 and M_3 with the same number of internal links l. They are all connected by one edge to the rest of the network M_0, and M_1 and M_3 are both connected to M_2 by one edge. The total number of links in the network is L. We assume they are communities within the modularity framework and in the weak sense according to Radicchi et al. (2) This implies that for $l^{out} = a * l$, we have $a \leq 2$ and $l < \frac{L}{4}$.
We look at three partitions of the network. In partition A, M_1, M_2 and M_3 are considered as separate modules. In B, M_1 and M_2 are one community, and in partition C all three modules are considered to be one single community.
Then the modularities for these partitions are:

$$Q_A = Q_0 + 2 * (\frac{l}{L} - (\frac{2l+2}{2L})^2) + \frac{l}{L} - (\frac{2l+3}{2L})^2 \quad [3]$$

$$Q_B = Q_0 + \frac{2l+1}{L} - (\frac{4l+4}{2L})^2 + \frac{l}{L} - (\frac{2l+2}{2L})^2 \quad [4]$$

$$Q_C = Q_0 + \frac{3l+2}{L} - (\frac{6l+7}{2L})^2 \quad [5]$$

Interest lies in the difference of the modularity between the two partitions:

$$\Delta Q_{AC} = Q_C - Q_A = \frac{2}{L} - (\frac{6l+7}{2L})^2 + \frac{2(2L+2)^2 - (2l+3)^2}{4L^2} = \frac{8L - 32l^2 - 80l - 50}{4L^2} \quad [6]$$

Because M_1, M_2 and M_3 are modules by construction, we expect partition A to have a higher modularity, i.e. that $\Delta Q_{AC} < 0$:

$$\frac{2}{L} < \frac{32l^2 + 80l + 50}{4L^2} = \frac{2(4l+5)^2}{4L^2} \Rightarrow l > \sqrt{\frac{L}{4}} - \frac{5}{4} \quad [7]$$

This inequality fails when $l < l_{AC}^{min} = \sqrt{\frac{L}{4}} - \frac{5}{4}$. Because for a fixed l, there is no constraint on the total number of edges L, modularity optimization may fail to detect the modules. The merged module from partition C has $3l + 2$ internal links. So if modularity optimization detects a module W with l_W links where $l_W < 3l_{AC}^{min} + 2 = 3\sqrt{\frac{L}{4}} - \frac{7}{4}$, then W may be a combination of the actual modules. Let's consider a network with $L = 900$. Then $l_{min} = \frac{55}{4} = 13.75$. I.e. when the modules have less than 13 links, they are under the resolution limit and modularity optimization would not detect them. A detected module W might be the combination of smaller modules when $l_W < \frac{173}{4} = 43.25$. Next follows the comparison of partition B with A.

$$\Delta Q_{AB} = Q_B - Q_A = \frac{1}{L} - \frac{(4l+4)^2 - (2l+3)^2}{(2L)^2} \quad [8]$$

The modules get detected when $\Delta Q_{AB} < 0$:

$$\frac{1}{L} < \frac{(4l+4)^2 - (2l+3)^2}{(2L)^2} \Rightarrow l > \frac{\sqrt{3L+1}}{3} - \frac{5}{6} \quad [9]$$

This inequality fails when $l < l_{AB}^{min} = \frac{\sqrt{3L+1}}{3} - \frac{5}{6}$ The merger of two modules contains $2l + 1$ links. Therefore a detected module W might actually consist of two of the actual modules when $l_W < 2l_{AB}^{min} + 1 = \frac{2\sqrt{3L+1}-2}{3}$. In the case of $L = 900$ this corresponds to $l_{AB}^{min} = \frac{\sqrt{2701}}{3} - \frac{5}{6} \approx 16.49$. I.e. when the modules have less than 16 links, modularity optimization would not detect them. A detected module W might be the combination of two smaller modules when $l_W < \frac{2\sqrt{2701}-2}{3} \approx 33.98$. The results show, that the resolution limit is higher, and therefore also the chance to detect the modules, when compared to the 2-cluster partition than to the 3-cluster one.

Consequences

For the illustration of this theoretical discussion, we analyzed an example of a real network. We used a undirected biological network of recombinant antigen genes from the human malaria parasite P. falciparum. The 307 nodes are var genes, and two genes are connected if they share a substring (on a highly variable region (HVR) in the DBLa domain of the var protein) whose length is statistically significant. The network contains a total of 2812 edges.

In order to detect the best community partition of the network we exploit Louvain Community Detection Algorithm. The procedure returned 8 communities corresponding to a very high modularity value ($Q_{max} \approx 0.636$). To tackle the issue of resolution limit,

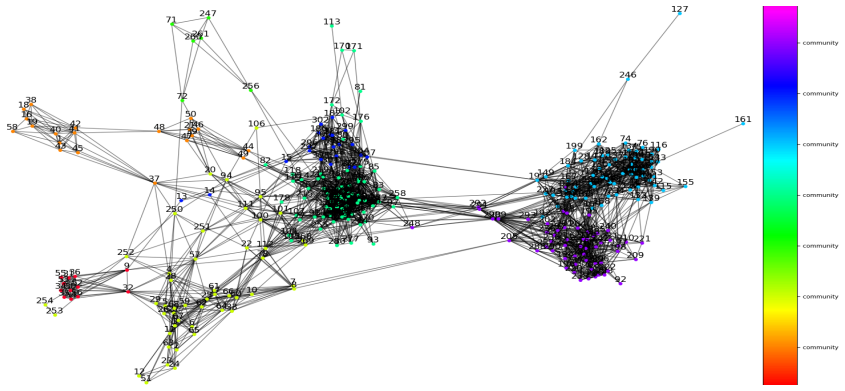

Fig. 2. The seven communities of malaria antigen genes that maximizes modularity.

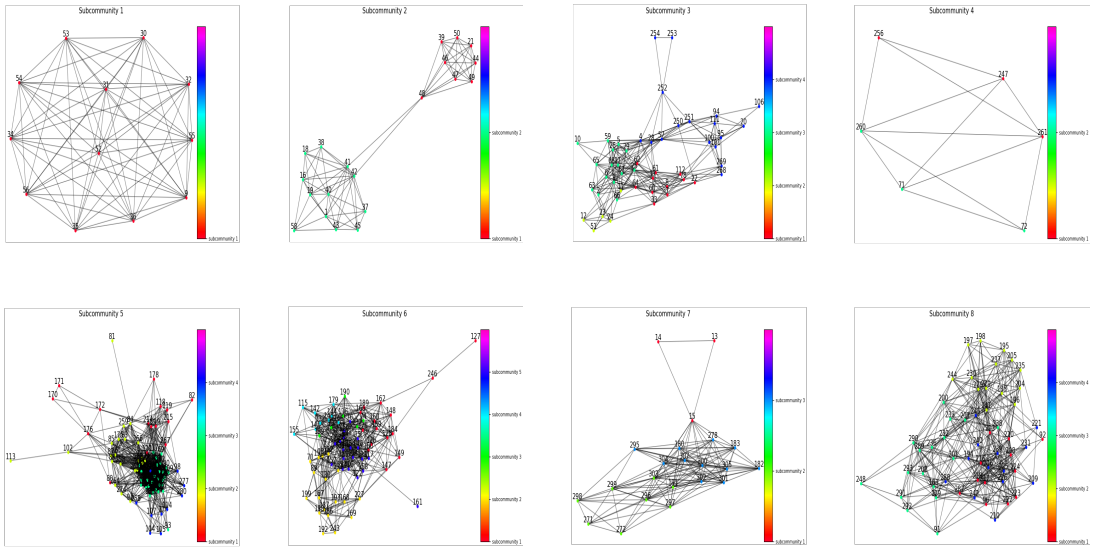

Fig. 3. Analysis of the finer structure of the 8 communities founded by modularity optimization. The indexes of the communities reflect those in Fig. 2

Table 1. Analysis of the composition and inner structure of the 8 moduli corresponding to the optimal modularity of the network

Modulus	n°nodes	n°edges	Q_{max}	n°submoduli
1	12	66	0.0	1
2	20	77	0.444	2
3	48	254	0.385	4
4	6	12	0.0	2
5	76	976	0.207	4
6	67	533	0.347	5
7	22	153	0.157	3
8	56	401	0.331	4

Enumeration of the communities correspond to the plot in Fig. 3

we constrain our analysis on each of the module and re-apply the Louvain Community Detection Algorithm. The underlying idea is to study the optimal modularity of each of the 8 modules to understand whether they are coherent groups or ensemble of smaller communities. This approach reveals the presence of a finer structure inside most of the detected communities. Worth-noticing are communities 2, 3, 6 and 8, whose Q_{max} is non-neglibile, as its values range from 0.331 to 0.444 (see Table 1). This proves that most of the modules founded by modularity optimization are in fact ensembles of smaller communities that the algorithm failed to identify.

Discussion

As already mentioned, the concept of the resolution limit in community detection is a significant factor in the realm of network analysis. It has profound implications that shape the way we understand complex networks. Firstly, it may result in the loss of important substructures of a network. This can be problematic when we're trying to understand complex networks where we have overlapping structures which can lead to a loss of crucial details. Additionally, the presence of a resolution limit can introduce bias into the results of community detection since potentially equally important sub-communities can remain hidden.

Quality functions that share mathematical similarities with modularity, wherein the partition's quality is determined by the sum of individual module qualities, likewise exhibit a resolution limit. This limitation can be explained through the same mathematical trade-off mentioned earlier. Therefore one must use a different way to define the quality of a partition:

An approach could be, as also mentioned in (1), to take the average quality of the modules, instead of the sum. This approach involves dividing the total quality by the number of modules, giving equal weight to each module.

$$\text{Average Quality} = \frac{\text{Total Quality}}{\text{Number of Modules}}$$

This helps prevent larger modules from dominating the evaluation. Additionally this also depends on the importance of considering the null model, which can either be global (influenced by the entire network) or local (influenced by the properties of the module alone). This choice can significantly impact the assessment of community structure. The decision on whether to use a global or local null model adds another layer of flexibility based on the specific characteristics and assumptions about the network under study. In Summary different quality functions may be suitable for different scenarios, and researchers should carefully consider the properties of the metric in relation to their study objectives.

Conclusion

We have explored the concept of the modularity resolution limit in a synthetic and a real network, namely the malaria antigen gene regulatory network. We find that the Louvain Community Detection Algorithm fails to detect some smaller communities present in the network, showing that the resolution limit is a problem in the analysis of real-world networks and that there is a need for methods that mitigate this problem. We propose ranking modules based on their average quality, according to some quality function that should be chosen by the researcher to be most appropriate for the network of interest. This way, partitions with larger modules will be less favorable, thereby reducing the resolution limit.

References

1. MB Santo Fortunato, Resolution limit in community detection. *PNAS* **104** (2006).
2. CFLVPD Radicchi F, Castellano C, The Comprehensive Tex Archive Network (CTAN). *Proc Natl Acad Sci USA* p. 101:2658–2663 (2004).