Modularity Measure of Networks With Overlapping Modules

Anna Lázár[1,2⋆], Dániel Ábel[1], and Tamás Vicsek[1,3]

[1] Department of Biological Physics, Eötvös University - Pázmány Péter stny. 1A, Budapest, Hungary H-1117
[2] Department of Mathematics, National University of Athens - Panepistimioupolis 15784 Athens, Greece
[3] Statistical and Biological Physics Research Group of HAS - Pázmány Péter stny. 1A, Budapest, Hungary H-1117

Abstract. In this paper we introduce a non-fuzzy measure which has been designed to rank the partitions of a network's nodes into overlapping communities. Such a measure can be useful for both quantifying clusters detected by various methods and during finding the overlapping community-structure by optimization methods. The theoretical problem referring to the separation of overlapping modules is discussed, and an example for possible applications is given as well.

1 Introduction

Networks – in the sense they are used throughout the present paper – are basically *graphs* describing various real-life complex systems. According to recent discoveries, they tend to have some interesting and rather unexpected common properties, such as scale-free degree distribution [1], strong disposition to form clusters (also called as communities or modules) or [2] exhibiting "small-world" property [3].

Since communities (groups of densely interconnected nodes) within these graphs often refer to the *functional units* of the corresponding complex systems, their exploration has been a fundamental issue. However, as an important result, these clusters turned out *not* to be separate, but rather overlapping, sharing many edges and nodes.

Because of the fundamental role clusters play in real-life networks, many algorithms have been proposed with the aim of uncovering the community-structure of a variety of networks. Earlier ones primarily detect disjoint clusters [9] [10], meanwhile some of the recent ones detect overlapping modules as well [2] [4] [5] [6] [7] [8].

At the same time, along with the development of the algorithms, the demand arose to define and measure somehow the *"suitability"* of the different partitions

⋆ This research was partially supported by the I.K.Y. Greek fellowship program and by the EU ERC COLLMOT grant.

provided by the various methods. Moreover, the fact that the concept of "cluster" is not specified enough (in the sense that it does not have a widely accepted definition) makes this problem even more ambiguous. However, although some of the proposed *measures* have become widely accepted and used (for example the so called "Q-modularity" proposed by Newman and Girvan in [9]), they are defined only for non-overlapping community structures.

Here we would like to note that *fuzzy* measures have been introduced to measure the "quality" of an overlapping community-structure, $\{c_1, \ldots, c_K\}$) [12] [13], but they share a common constraint: every i node has a "belonging factor" $0 \leq \alpha_{i,c_r} \leq 1$ which expresses how *strongly* node i belongs to the rth cluster c_r. The requirement is that $\sum_{r=1}^{K} \alpha_{i,c_r} = 1$ for all i node belonging to the graph, K denoting the number of clusters.
In other words, *none of the nodes can belong to more than one community "strongly"* (and, primarily, not "fully"). Recalling social networks, this means that if a person belongs – let's say – to her/his family fully (or "strongly"), then she/he can not belong to other communities, like working place, sport club, etc, only very "weakly", or nohow. We believe that this condition is often un-realistic in real-life cases, so our goal has been to define a measure without the above requirement.

In brief, the purpose of the present paper is to define a simple but well-usable non-fuzzy measure which, on the one hand, quantifies cluster-structures found by various methods on connected networks, and on the other hand, can be used to detect (overlapping) communities as well by directly optimizing it. For being well-usable, we expect from the measure to take values between -1 and and 1, where a higher positive value corresponds to a better clustering. The zero value expresses random-like network-clustering, that is, when the clusters are created randomly.

2 The proposed measure

Since the notion of "cluster" is not well-defined, many different measures can be conceived, according to the different "intuitive" characteristics: average path-length among nodes, betweenness, etc. However, the most commonly used ones exploit the expectation that a cluster should be *"dense"* – or, as it is often formulated: modules are expected to have relatively more connections within themselves than among each other [9] [11] [12]. Using the above expectation (denseness) *and* allowing overlapping community-structure leads to the result that *separate edges* will be returned as optimal community-structure – since these are the most dense subgraphs, see fig. 1 a. (This happens for example if one tries to apply Newman's Q-modularity directly onto structures where overlapping is enabled.)

According to our experiments, *none* of the "intuitive approaches" is enough to create a suitable measure *alone*, because they result in "degenerated struc-

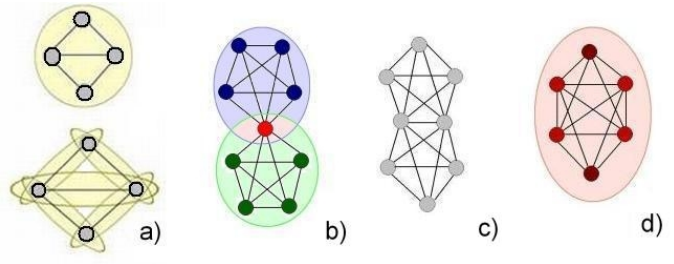

Fig. 1. ***a)*** Measure based on the modules *density* will be optimal if all the edges constitute a separate cluster. ***b)-d)***: The question when to handle a (sub)graph as one community and when as more, is non-trivial, because "intuition" gives different answers to different people. At the same time, most of us would agree on separating two 5-cliques overlapping in one single node ***(b)***, but handling them as one community, if they share 4 nodes ***(d)***. Cases between ***(c)*** are a matter of "taste".

tures" to be optimal ones, similar to the one seen above. On the other hand, *combinations* of approaches can handle this phenomenon.
We have obtained good results by utilizing the following expectations: (1) the edges of a given node should primarily go inward its cluster(s) and should not go outward, and (2): clusters should be dense. The first criterion shows how "justifiable" it is to assign the node $i(\in c_r)$ to the rth cluster c_r: it is the difference between the *inward* going edges ($\sum_{j\in c_r, i\neq j} a_{ij}$) and the *outward* going edges ($\sum_{j\notin c_r} a_{ij}$), divided by the d_i degree of node i. Put it together, we get that every i node contributes to the rth cluster to which it belongs to with the following value:

$$\frac{\sum\limits_{j\in c_r, i\neq j} a_{ij} - \sum\limits_{j\notin c_r} a_{ij}}{d_i} \tag{1}$$

where $a_{i,j}$ denotes the proper element of the adjacency matrix defining the network, interpreted as usually, that is,

$$a_{ij} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ are connected,} \\ 0 \text{ if not} \end{cases} \tag{2}$$

The more edges go inward and the less edges go outward the cluster, the more the above ratio converges to 1. If more edges go outward than inward, the expression is negative, and if all of them go outward, the result is -1. Due to the overlapping areas, a node can contribute with positive values to more than one clusters.

To avoid community-structures having only a few communities with very high $M_{c_r}^{ov}$ values, we add the criterion that all nodes have to belong to at least one module. (A trivial solution for that is, to put all the left-out nodes into a separate cluster at the end. We have obtained our results like this too.) Also, the appearance of many almost, or entirely overlapping communities is avoidable by

dividing the above expression by the number of clusters i belongs to, denoted by s_i. Thus the rth cluster, c_r will contribute to the final result M^{ov} with:

$$M^{ov}_{c_r} = \frac{1}{n_{c_r}} \sum_{i \in c_r} \frac{\sum\limits_{j \in c_r, i \neq j} a_{ij} - \sum\limits_{j \notin c_r} a_{ij}}{d_i \cdot s_i} \cdot \frac{n^e_{c_r}}{\binom{n_{c_r}}{2}} \tag{3}$$

where n_{c_r} is the number of nodes and $n^e_{c_r}$ is the number of edges that the rth cluster c_r contains, respectively.

The *density* of a module – which was our "second requirement" – is straightforward to interpret as $\frac{n^e_{c_r}}{\binom{n_{c_r}}{2}}$. This expression gives 1 if the rth module c_r (which is a (sub)graph) contains all its possible edges, and 0 if it does not have any of them. Since the first factor ranges between -1 and 1, the second factor between 0 and 1, the whole expression varies between -1 and 1.

This remains true for the final measure M^{ov} as well, which is the average of the $M^{ov}_{c_r}$ module-values:
$M^{ov} = \frac{1}{K} \sum\limits_{r=1}^{K} M^{ov}_{c_r}$, that is,

$$M^{ov} = \frac{1}{K} \sum_{r=1}^{K} \left[\frac{\sum\limits_{i \in c_r} \frac{\sum\limits_{j \in c_r, i \neq j} a_{ij} - \sum\limits_{j \notin c_r} a_{ij}}{d_i \cdot s_i}}{n_{c_r}} \cdot \frac{n^e_{c_r}}{\binom{n_{c_r}}{2}} \right] \tag{4}$$

Since the density of clusters containing one single node (when $n_{c_r} = 1$) is not defined (because $\binom{1}{2}$ is not defined), we simply set their $M^{ov}_{c_r}$ modularity value to zero. (Isolated nodes (when $d = 0$) can not appear, since the network assumed to be connected.)

Here we would like to note that handling the unclustered nodes (nodes that do not belong to any of the modules) is possible in many ways. We have chosen to put them into a separate community, but some kind of *weighting* is also conceivable, when the weight is in inverse proportion to the number of the unclustered nodes (the more nodes are clustered, the higher the final score is). Furthermore, one can consider the *weighting* of the clusters according to their sizes as well.

3 One cluster or more clusters? – When to separate?

This question is highly non-trivial, because it is – up to a great extent – simply a matter of "intuition" or taste, being different from person to person. For example most of us would agree on separating two 5-cliques overlapping in one single node, but handling them as one community, if they share 4 nodes (see fig. 1). But what is the case, if they share two or three nodes?

Figure 2 a) and b) describes how the introduced measure, M^{ov} behaves with respect to the above question. Given a complete-graph with $n_2 = 50$ nodes and a smaller one with n_1 nodes ($n_1 \in \{1 \dots 50\}$, also complete-graph). These two graphs overlap in o nodes, where $o \in \{1 \dots n_1\}$ (Fig. 2 b)). On sub-picture a), the horizontal axis shows the size of the smaller graph, n_1, while the vertical axis shows the number of the overlapping nodes (o) between the two graphs. Two regions show up: the lower region covers the $o-n_1$ parameter-pairs by which M^{ov} gives higher score if the two graphs are handled as separate communities, while the upper one covers those n_1-o pairs, which give higher score, if the overlapping graphs are handled as one module. One extreme is when the overlap is 0 (the two graphs do not share any nodes, horizontal axis) – which obviously falls in the lower, "separate"-region. The other end-value is when they share all the n_1 nodes, that is, the smaller graph (the n_1-clique) is a real sub-graph, a part of the bigger complete-graph – this case is represented by the diagonal line starting from the pole.

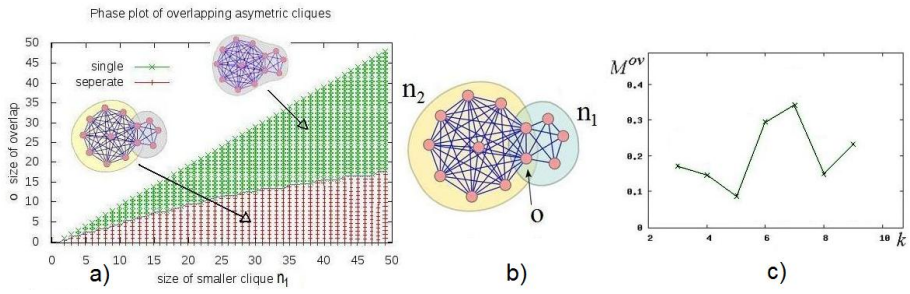

Fig. 2. ***a)-b)*** Given a complete-graph with $n_2 = 50$ nodes and a smaller one with n_1 nodes ($n_1 \in \{1 \dots 50\}$, also complete-graph; n_1 is shown on the horizontal axis). These two graphs overlap in o nodes (vertical axis), where $o \in \{1 \dots n_1\}$. The $n_1 - o$ parameter-pairs generate two dissevering regions: the upper one is where the introduced measure, M^{ov} gives higher score if the two graphs are handled as one module, while the lower one covers those $n_1 - o$ pairs, which give higher score if the graphs make up separate communities. ***c)*** The M^{ov} scores as a function of the k "tuning-parameter" belonging to the CFinder algorithm, for the protein-protein interaction network. The suggested k-value is where the curve reaches its maximum, that is, at $k = 7$.

4 An application

CFinder, an algorithm designed to uncover the overlapping community-structure of networks [2], has a "tuning-parameter" (k) which determines the *cohesiveness* of the revealed modules: the higher the parameter k, the smaller, the more disintegrated, but at the same time the more cohesive are the detected communities. Theoretically k can be any positive integer starting from 3, but in practice it is

usually smaller than ten. (If $k = 2$, CFinder detects the connected subgraphs.) The proper value of k depends on the network. In the following we define the most proper k for a real-life network using the introduced measure, M^{ov}.
Figure 2 c) depicts the M^{ov} scores as a function of the k parameter for a network describing the protein-protein interactions in S. cerevisiae (see details in [17]). As it can be seen, the curve reaches its maximum at $k = 7$, that is, using the CFinder algorithm, the best (overlapping) community structure is revealed by setting the k tuning parameter to 7.

We would also like to highlight, that the best possible M^{ov} score very strongly depends on the network itself.

References

1. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Reviews of Modern Physics, 74, 47–97 (2002)
2. Palla, G., Derényi, I., Farkas, I., Vicsek T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814–818 (2005)
3. Watts, D. J., Strogatz, S. H.: Collective dynamics of 'small-world' networks. Nature, 393, 440–442 (1998)
4. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex systems. New J. Phys. 11, 033015 (2009)
5. Shen, H., Cheng, X., Cai, K., Hu, M.: Detect overlapping and hierarchical community structure in networks. Physica A. 388, 1706–1712 (2009)
6. Ahn, Y., Bagrow, J. P., Lehmann, S.: Link communities reveal multi-scale complexity in networks. arXiv:0903.3178v2.
7. Li, D., Leyva, I., Almendral, J. A., Sendina-Nadal, I., Buldu, J. M., Havlin, S., Boccaletti, S.: Synchronization interfaces and overlapping communities in complex networks. Physical Review Letters, 101, 168701 (2008)
8. Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., Vicsek T.: CFinder: Locating cliques and overlapping modules in biological networks. Bioinformatics, 22, 1021–1023 (2006)
9. Newman, M.E.J., Girvan M.: Finding and evaluating community structure in networks. Phys. Rev. E. 69, 026113 (2004)
10. Newman, M.E.J.: Modularity and community structure in networks. Proc. of the Nat. Academy of Sciences of the USA (PNAS). 103, 8577–8582 (2006)
11. Leicht, E.A., Newman, M.E.J.: Community structure in directed networks. Phys. Rev. Lett., 100, 118703 (2008)
12. Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri M.: Extending the definition of modularity to directed graphs with overlapping communities. J. Stat. Mech. P03024 (2009)
13. Nepusz, T., Petróczi, A., Négyessy, L., Bazsó, F.: Fuzzy communities and the concept of bridgeness in complex networks. Physical Review E, 77, 016107 (2008)
14. Scott, J.: Social Network Analysis: A Handbook. Sage Publications, London, (2000)
15. Everitt, B. S.: Cluster Analysis. Edward Arnold, London (1993)
16. Newman, M.E.J.: Detecting community structure in networks. Eur. Phys. J. B. 38, 321–330 (2004)
17. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., Eisenberg, D.: DIP: the Database of Interacting Proteins. Nucleic Acids Res. 28(1), 289–291 (2000)