A Letters Journal Exploring the Frontiers of Physics

April 2010

EPL, **90** (2010) 18001 www.epljournal.org
doi: 10.1209/0295-5075/90/18001

Modularity measure of networks with overlapping communities

A. Lázár[1(a)], D. Ábel[1] and T. Vicsek[1,2]

[1] *Department of Biological Physics, Eötvös University - Pázmány Péter stny. 1A, Budapest, H-1117 Hungary, EU*
[2] *Statistical and Biological Physics Research Group of HAS - Pázmány Péter stny. 1A, Budapest, H-1117 Hungary, EU*

received 14 October 2009; accepted in final form 23 March 2010
published online 27 April 2010

PACS 89.75.Hc – Networks and genealogical trees
PACS 89.75.Fb – Structures and organization in complex systems
PACS 05.90.+m – Other topics in statistical physics, thermodynamics, and nonlinear dynamical systems

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.

Introduction. – Networks —in the sense they are used throughout the present paper— are basically *graphs* describing real-life complex systems taken from the most different scientific areas, but primarily from biology, economy and sociology. According to recent discoveries, real-life networks tend to have some interesting and rather unexpected common properties, such as scale-free degree distribution, strong disposition to form clusters (also called as communities or modules) or having the so-called "small-world" property [1–3].

Communities (groups of densely interconnected nodes) within these graphs often refer to the *functional units* of the corresponding complex systems, thus their exploration has been a fundamental issue in the study of networks. 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 [4,5], meanwhile some of the recent ones detect overlapping modules as well [2,6–10]. In a very interesting recent paper, the authors propose to interpret communities as groups of *links* instead of groups of nodes, since edges often exist because of *one* dominant reason, even if the nodes they connect belong to different communities [8]. Another neat solution has been introduced in [9], in which a practical method is suggested to define the overlap by using synchronized oscillators. In this paper the authors study the interfaces appearing in complex networks between the clusters of phase oscillators. The algorithm introduced in [7] uncovers the overlapping and hierarchical properties of networks based on sets of maximal cliques.

Along with the development of the algorithms, the demand also arose to define and measure somehow the *"suitability"* of the different partitions 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 [4]), they are defined only for non-overlapping community structures.

Here we would like to note that *fuzzy* measures have been introduced with the same ambition (namely to measure the "quality" of an overlapping community structure, $\{c_1, \ldots, c_K\}$) [11,12] but they share a common constraint: every i node has a "belonging factor" $0 \leqslant \alpha_{i,c_r} \leqslant 1$ which expresses how *strongly* node i belongs to the r-th cluster c_r. The requirement is that

$$\sum_{r=1}^{K} \alpha_{i,c_r} = 1 \tag{1}$$

for all i belonging to the graph, K denoting the number of clusters.

(a) E-mail: lanna@hal.elte.hu

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 us say— to her/his family fully (or "strongly"), then she/he cannot belong to other communities, like working place, sport club, etc, only very "weakly", or nohow. We believe that this condition is often unrealistic 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 the measure to take values between -1 and 1, where a higher positive value corresponds to a better clustering. We also expect our measure to give results near to zero for random-like cluster partitions, that is, when clusters are chosen randomly for an arbitrary network. To test this expectation, we have defined 10000 random cluster partitions for an arbitrary graph, which in this case has been the one depicted in fig. 6(a). The process for defining a random clustering has been the following: first we have generated a random number taking values between 1 and the number of nodes, using a function which generates an integer value from the uniform distribution on the given set: this value determined the *number of clusters*. Then, for all of these clusters, *the number of the nodes* have been defined using the very same function, that is, how many nodes the several clusters should include. And finally, the nodes have been chosen randomly as well, for all the clusters. We have calculated the score of our measure for all these cluster partitions and defined their average, which has been of the order of magnitude 10^{-3}, that is, near to zero.

The proposed measure. – As mentioned above, the notion of "cluster" is not well defined: there are many approaches based on different "intuitive" characteristics of a community, such as its denseness, the average path-length among its nodes, the number of edges going in and out of a given module, the betweenness among nodes belonging to different communities, etc. [13–15]. Although theoretically, measures could be constructed based on any of the above characteristics, in practice, 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 and than among each other [4,11,16]. Using the above expectation (clusters should be dense) *and* allowing overlapping community structure leads to the result that *separate edges* will be returned as optimal community structure —since these are the densest 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.)

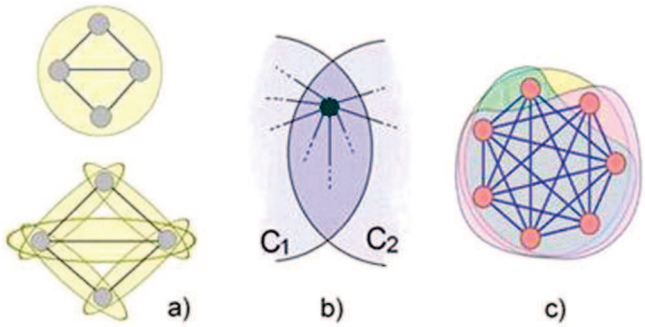

Fig. 1: (Colour on-line) (a) Measure based on the modules *density* will be optimal if all the edges constitute a separate cluster. (b) An overlapping node that belongs to both the c_1 and c_2 communities. It contributes with positive values for both clusters. (c) The appearance of many similar or almost-the-same overlapping communities.

According to our experiments, *none* of the "intuitive approaches" is enough to create a suitable measure *alone*, because they result in "degenerated structures" 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 r-th 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 r-th cluster to which it belongs to with the following value:

$$\frac{\sum\limits_{j\in c_r, i\neq j} a_{ij} - \sum\limits_{j\notin c_r} a_{ij}}{d_i}, \tag{2}$$

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

$$a_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are connected,} \\ 0, & \text{if not.} \end{cases} \tag{3}$$

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. Since a node can contribute with positive values to more than one clusters —due to the overlapping areas, see fig. 1(b)— the whole network's modularity value is higher if a node like that belongs to both modules.

In order to prevent the appearance of many similar or almost-the-same overlapping communities (as can be seen on fig. 1(c)) the above expression is divided by the

number of clusters i belongs to, denoted by s_i. Thus the r-th cluster, c_r will contribute to the final result M^{ov} with

$$M^{ov}_{c_r} = \frac{1}{n_{c_r}} \sum_{i \in c_r} \frac{\sum\limits_{j \in c_r, i \neq j} a_{ij} - \sum\limits_{j \notin c_r} a_{ij}}{d_i \cdot s_i} \cdot \frac{n^e_{c_r}}{\binom{n_{c_r}}{2}}, \qquad (4)$$

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

The *density* of a module —which was our "second requirement"— is straightforward to be interpreted as $\frac{n^e_{c_r}}{\binom{n_{c_r}}{2}}$. This expression gives 1 if the r-th 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:

$$M^{ov} = \frac{1}{K} \sum_{r=1}^{K} M^{ov}_{c_r},$$

that is

$$M^{ov} = \frac{1}{K} \sum_{r=1}^{K} \left[\frac{\sum\limits_{i \in c_r} \frac{\sum\limits_{j \in c_r, i \neq j} a_{ij} - \sum\limits_{j \notin c_r} a_{ij}}{d_i \cdot s_i}}{n_{c_r}} \cdot \frac{n^e_{c_r}}{\binom{n_{c_r}}{2}} \right]. \qquad (5)$$

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 $d = 0$) cannot appear, since the network assumed to be connected.)

To avoid community structures having only a few communities with very high $M^{ov}_{c_r}$ 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.)

On the other hand, 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.

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. 2). But what is the case, if they share two or three nodes?

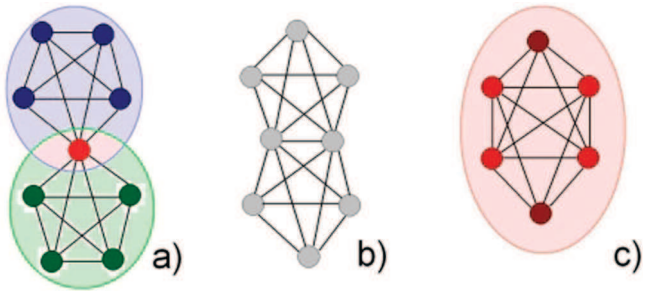

Fig. 2: (Colour on-line) The question when to handle a (sub)graph as one community and when as more, is non-trivial, because "intuition" gives different answers to different people. At the same time, most of us would agree on separating two 5-cliques overlapping in one single node (a), but handling them as one community, if they share 4 nodes (c). Cases between (b) are a matter of "taste".

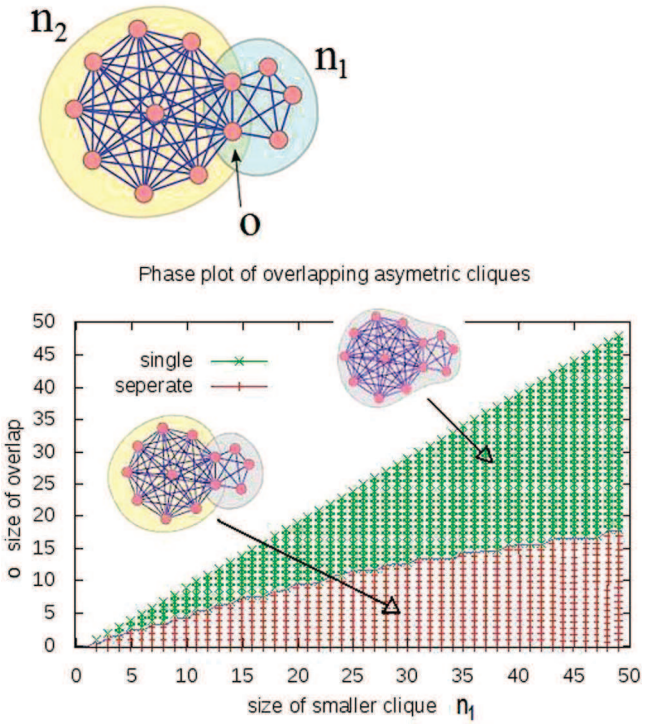

Fig. 3: (Colour on-line) Given a complete graph with $n_2 = 50$ nodes and a smaller one with n_1 nodes ($n_1 \in \{1 \dots 50\}$, also complete graph; n_1 is shown on the horizontal axis). These two graphs overlap in o nodes (vertical axis), where $o \in \{1 \dots n_1\}$. The n_1-o parameter pairs generate 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.

Figure 3 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 $o \in \{1, \dots, n_1\}$. 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 a 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.

As the overlap between the two complete cliques grows, assuming a single community gives better M^{ov} scores. The border line between these two areas depends on the formulation of the function used for scoring the community structure. Judging whether the border line resulting from a given definition is good or bad is, of course, highly subjective. It is quite reasonable to assume that above a certain overlap size, the single-community structure should get a higher score, while below that two communities should be optimal. One feature of the plot that can be objectively judged is the fact that such a transition occurs for even small n_1, *i.e.* our proposed method prefers a small clique to be a separate community (provided that the overlap between them is not too large), even if there is a large size difference between the two cliques.

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. This is a result of the method, which exploits the observation, that a typical community consists of several complete subgraphs that tend to share many of their nodes. The algorithm uncovers those modules which form so-called "k-clique communities", that is, unions of k-cliques that can be reached from each other through a series of adjacent k-cliques.

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, that is, those modules which are unions of 2-cliques (which are edges) and can be reached from each other through a series of adjacent edges.) The proper value of k depends on the network. In the following we define the most proper k for some real-life networks using the introduced measure, M^{ov}.

Figure 4 depicts the M^{ov} scores as a function of the k-parameter for three real-life networks: 1) word association, 2) protein interaction, and 3) cond-mat publication.

The nodes of the first graph, "word association", are words which are linked if the people in a survey associated them with each other [17]. (Originally it is a

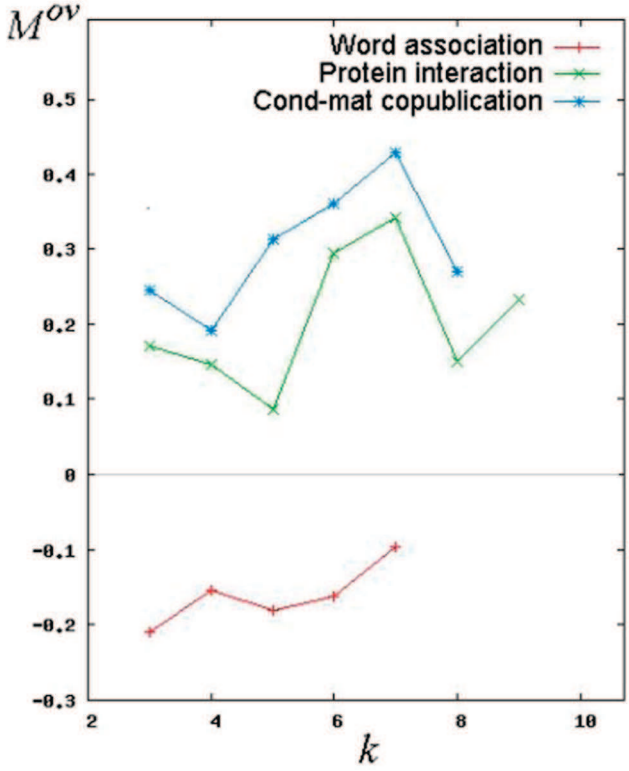

Fig. 4: (Colour on-line) The M^{ov} scores as a function of the k "tuning-parameter" belonging to the CFinder algorithm, for three real-life networks: 1) cond-mat publication (topmost curve) 2) protein interaction and 3) word association (bottom-most curve). The suggested k-values are those where the curves reach their maximum.

weighted, directed graph, where the weight of an edge indicates the frequency that the people associated the end point of the link with its start point, but here we have used a simplified —undirected, unweighted— form of it.) The "protein interaction" network describes the protein-protein interactions in *S. cerevisiae* (see details in [18]), and finally, the "cond-mat publication" network describes co-authorships among mathematicians, obtained from the Los Alamos cond-mat archive [19]. (Originally this is a weighted graph as well, where the weights are proportional to the number of common works, but, here too, we have used a simplified, unweighted version of the graph, in which the edges have been eliminated under a certain threshold-weight. See more details in [20].)

As can be seen in fig. 4, in the case of the protein-interaction network and the cond-mat publication, both curves reach their maximum at $k=7$, which is their optimum value for k.

At the same time, randomizing these networks by randomly distributing the edges results in M^{ov} scores much lower than for the corresponding real network: -0.06, -0.3 and -0.57 at $k=3$ for the co-publication, protein interaction and word association networks respectively, with no communities found for higher values of k.

The word association network displays a very interesting behavior: the whole curve is in the negative region.

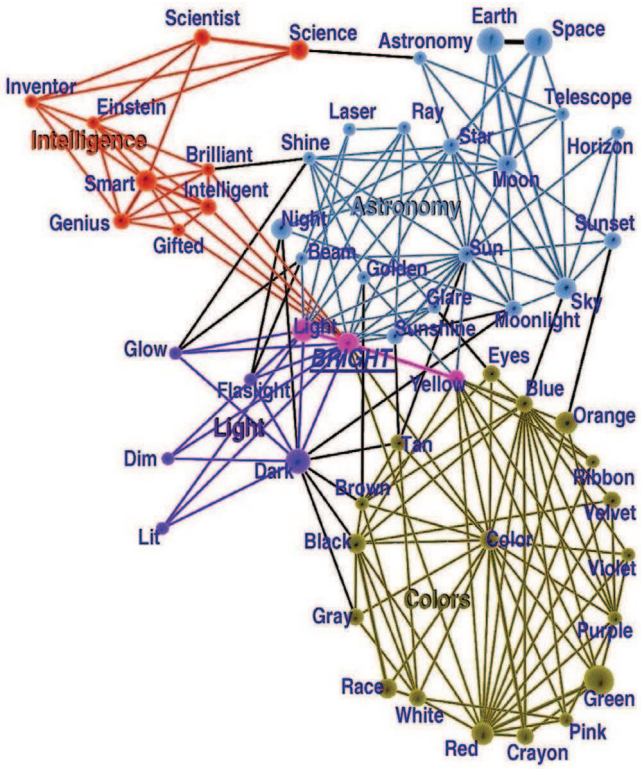

Fig. 5: (Colour on-line) The densely interconnected, frequent associations of the word "bright". This sub-graph is a typical segment of the "word association network" [17] which has been used to provide the bottommost curve in fig. 4 ("word association curve") which exhibits an interesting behavior by being entirely in the negative region.

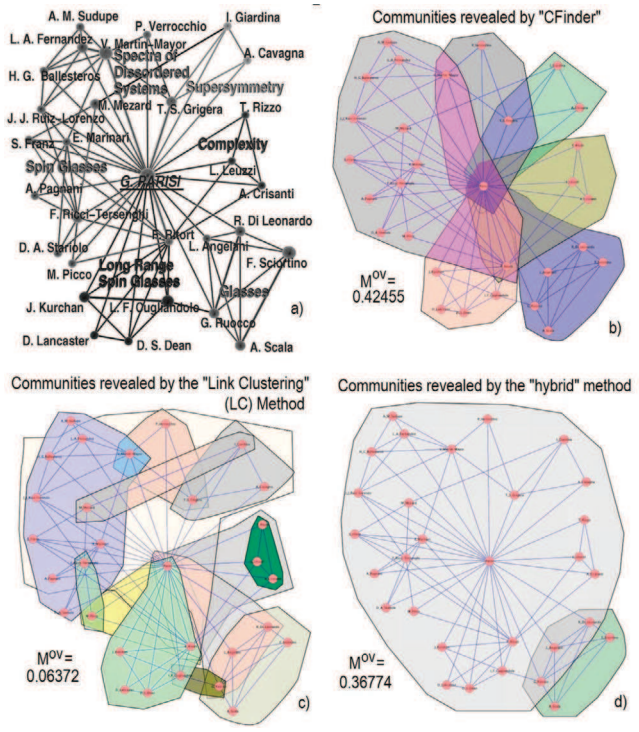

Fig. 6: (Colour on-line) (a) The network we have used for comparing the M^{ov} values belonging to different partitionings: the densely interconnected parts of the co-authorship relations of G. Parisi based on the Los Alamos Condensed Matter archive [19]. (b) The clustering revealed by the CFinder algorithm; $M^{ov}_{CFinder} = 0.42455$. (c) Cluster partitions determined by the Link Clustering Method; $M^{ov}_{LC} = 0.06372$. (d) Partitioning obtained by using the "hybrid" method, for which $M^{ov}_{Hybrid} = 0.36774$.

This is most probably due to the fact that this graph contains many words with several meanings, *e.g.*, the word "bright", which —according to the survey— is often associated with words having alternative meanings, like "smart", "light", "dark", "sun", etc. (see fig. 5) Accordingly, in a graph like this, if slightly overlapping modules arise around the different meanings of a word, and if between the nodes of these otherwise separate modules there are relatively many edges (associations) a negative numerator in M^{ov} results.

Figure 5 illustrates the densely interconnected, frequent associations of the word "bright". This sub-graph is a typical segment of the "word association network" which has been used to provide the bottommost curve in fig. 4 ("word association curve") by identifying the communities with the "CFinder" algorithm using several k values. As it can be seen in fig. 5, this graph has a special feature, namely that —apart from containing nodes with similar meanings— it has nodes with *opposite* meanings as well, which are often connected: for example "bright" has a link to "dark". Thus nodes belonging to contradictory categories (clusters) can be easily connected.

We have investigated the behavior of this network with a recent, neat algorithm as well —namely with the Link Clustering (LC) Method [8]— and have found that the M^{ov} values belonging to the cluster partitions revealed by the LC algorithm, are negative as well. Since LC detects reasonable clusterings in general (with positive M^{ov} values), it is well-founded to assume that it is indeed a feature of the network that calls forth negative M^{ov} values.

Cluster partitions and M^{ov} values —a comparison. – In order to compare the M^{ov} values belonging to various partitionings on the same network, we have determined the cluster partitions using two different methods and a "hybrid" one, and calculated the corresponding M^{ov} values for each of them, respectively. The graph we have used is a small, real-life network showing the densely interconnected parts of the co-authorship relations of G. Parisi, based on the database [19] (see fig. 6(a)).

The clustering methods we have used are the following:

- CFinder [2];
- Link Clustering (LC) Method [8];
- a "hybrid" method which refers to the following process: for determining the optimal partitioning, the previous method (that is the Link Clustering

algorithm) builds a dendrogram during its run, and defines the optimal clustering by using a function called "Partition Density": the dendrogram is cut where this function is maximal. In our comparison experiments, instead of cutting this dendrogram where the Partition Density function has been maximal, we have cut it where the M^{ov} was maximal.

Figure 6(b) depicts the clustering revealed by the CFinder algorithm. The corresponding M^{ov} value is $M^{ov}_{CFinder} = 0.42455$. Figure 6(c) shows the cluster-partitions returned by the Link Clustering Method; the corresponding M^{ov} value is $M^{ov}_{LC} = 0.06372$. Finally, fig. 6(d) depicts the cluster partitions determined by the "hybrid" method, for which $M^{ov}_{Hybrid} = 0.36774$.

In our example network (fig. 6(a)) some of the overlapping nodes are those which are at the same time central to their clusters as well (for example the node representing G. Parisi). These kinds of cluster structures are a special challenge for the algorithms detecting cluster partitions. Figure 6 exemplifies that those clusterings possess higher M^{ov} scores, which appear to be better in an "intuitive" way.

* * *

This research was partially supported by the grant EU ERC COLLMOT.

REFERENCES

[1] Albert R. and Barabási A.-L., *Rev. Mod. Phys.*, **74** (2002) 47.
[2] Palla G., Derényi I., Farkas I. and Vicsek T., *Nature*, **435** (2005) 814.
[3] Watts D. J. and Strogatz S. H., *Nature*, **393** (1998) 440.
[4] Newman M. E. J. and Girvan M., *Phys. Rev. E.*, **69** (2004) 026113.
[5] Newman M. E. J., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 8577.
[6] Lancichinetti A., Fortunato S. and Kertész J., *New J. Phys.*, **11** (2009) 033015.
[7] Shen H., Cheng X., Cai K. and Hu M., *Physica A*, **388** (2009) 1706.
[8] Ahn Y., Bagrow J. P. and Lehmann S., *Link communities reveal multi-scale complexity in networks*, arXiv:0903.3178v2.
[9] Li D., Leyva I., Almendral J. A., Sendina-Nadal I., Buldu J. M., Havlin S. and Boccaletti S., *Phys. Rev. Lett.*, **101** (2008) 168701.
[10] Adamcsek B., Palla G., Farkas I. J., Derényi I. and Vicsek T., *Bioinformatics*, **22** (2006) 1021.
[11] Nicosia V., Mangioni G., Carchiolo V. and Malgeri M., *J. Stat. Mech.* (2009) P03024.
[12] Nepusz T., Petróczi A., Négyessy L. and Bazsó F., *Phys. Rev. E*, **77** (2008) 016107.
[13] Scott J., *Social Network Analysis: A Handbook* (Sage Publications, London) 2000.
[14] Everitt B. S., *Cluster Analysis* (Edward Arnold, London) 1993.
[15] Newman M. E. J., *Eur. Phys. J. B*, **38** (2004) 321.
[16] Leicht E. A. and Newman M. E. J., *Phys. Rev. Lett.*, **100** (2008) 118703.
[17] Nelson D. L., McEvoy C. L. and Schreiber T. A., The University of South Florida word association, rhyme, and word fragment norms, `http://web.usf.edu/FreeAssociation/` (1998).
[18] Xenarios I., Rice D. W., Salwinski L., Baron M. K., Marcotte E. M. and Eisenberg D., *Nucleic Acids Res.*, **28** (2000) 289.
[19] `http://arxiv.org/archive/cond-mat`.
[20] Warner S., *Libr. Hi Tech.*, **21** (2003) 151.