Journal of Complex Networks (2023) **2**, Advance Access Publication on 17 April 2023
https://doi.org/10.1093/comnet/cnad012

Targeted Community Merging provides an efficient comparison between collaboration clusters and departmental partitions

F. J. Bauza[†]

Department of Theoretical Physics, University of Zaragoza, 50009 Zaragoza, Spain
[†]Corresponding author. Email: fbm.prof@hotmail.com

G. Ruiz-Manzanares

Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018 Zaragoza, Spain

J. Gómez-Gardeñes

Department of Condensed Matter Physics, University of Zaragoza, 50009 Zaragoza, Spain

A. Tarancón and D. Íñiguez

Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018 Zaragoza, Spain

[Received on 8 February 2023; editorial decision on 29 March 2023; accepted on 5 April 2023]

Community detection theory is vital for the structural analysis of many types of complex networks, especially for human-like collaboration networks. In this work, we present a new community detection algorithm, the Targeted Community Merging algorithm, based on the well-known Girvan–Newman algorithm, which allows obtaining community partitions with high values of modularity and a small number of communities. We then perform an analysis and comparison between the departmental and community structure of scientific collaboration networks within the University of Zaragoza. Thus, we draw valuable conclusions from the inter- and intra-departmental collaboration structure that could be useful to take decisions on an eventual departmental restructuring.

Keywords: complex networks; community structure; scientific collaboration networks; similarity indices.

1. Introduction

Network theory has been widely used in modelling of all types of complex systems [1–3]. Complex networks provide us with a versatile and easy to handle picture of the structure of interactions among agents of a complex system, which allows studying both its structural and functional properties. The analysis of structural properties of a complex system has been the focus of countless works in recent years, either because of the importance of the structure of interactions itself (network robustness and navigation properties, network dimension, community and k-core structures, etc.) [4–13], or because of the crucial role that the interaction backbone has proven to play in the performance of dynamic models of different types (epidemics, opinions, game theory and synchronization) [14–20].

One of the fundamental pillars of structural network analysis is the detection and analysis of communities [21–23]. Within a network, a community is understood as a group of nodes among which interactions are more frequent (or of greater weight) than would be expected if interactions were completely random [3]. The detection and analysis of this type of groups give us relevant information on the characteristics of the structure of interactions at a mesoscopic scale, halfway between the global and

local scale. In the last two decades, a large number of results have been obtained in the study of the detection and analysis of communities in networks, with the corresponding development of a great variety of community detection algorithms. There exist two main branches in which most of these algorithms are included. The first one consists of algorithms [24–27] maximizing a measure of the goodness of the community partition called modularity [28]. Out of this wide range of algorithms, it could be pointed out the iterative Girvan–Newman algorithm (GN) [29], being one of the best known and most widely used. The second branch consists of algorithms that are framed within a recently developed model for graph generation, called Stochastic Block Models [30–33].

One of the main applications of complex networks is in the field of collaboration networks. These networks attempt to model complex systems of different nature (socioeconomic, industrial, biomedical or research-oriented) in which one or many common goals are to be achieved through the collaboration of agents, usually distributed among different collaboration groups [34–36]. The analysis of research collaboration networks [37–42] has a special interest, aligned with the field of Science of Science [43], since community detection analysis can be used straightforward to identify research alliances within agents and disciplines with particular scientific advancements.

In this work, we present a new modularity maximization community detection algorithm called the Targeted Community Merging (TCM), based on the well-known GN algorithm. Like the GN algorithm, TCM is an iterative algorithm which allows to obtain different partitions with a different number of communities per partition; however, the main advantage of this new algorithm is that it allows to reduce considerably the number of communities per partition with hardly any damage to modularity. This feature is really useful when comparing the optimal community partition with some native partition in real networks, especially for human-kind collaboration networks, where the number of groups in the native partition is usually small. Here, we apply this new algorithm to obtain the optimal community partition in different real-world researchers collaboration networks, for several macro-areas of the University of Zaragoza.

Some universities are considering a modification of their departmental structure, generally established many years ago and based on teaching criteria, in order to gain efficiency both in economic and scientific terms. The point of obtaining these optimal partitions is that we could propose a restructuring of the actual departmental partition of each macro-area, that would be more based on enhancing collaboration than on branches of knowledge. Besides, we compare the community partitions with the partitions into departments, so as we can unveil the insights of the inter- and intra-departmental collaborative structure. In addition, we make use of some part of the knowledge of similarity of sets, specifically the Rand and Wallace similarity indices [44, 45], both for selecting the most appropriate community partition and as part of the quantitative comparison between the community partition and the departmental one.

The article is organized as follows. First, in Section 2, we describe the data used and the assumptions considered to construct the researchers collaboration networks (Section 2.1), we present the TCM algorithm (Section 2.2), and we introduce the similarity indices used to compare the community and departmental partitions (Section 2.3). In Section 3, we show the main results of the out-performance of the TCM algorithm over the usual GN algorithm (Section 3.1) and the main properties of the inter- and intra-departmental collaborative structure of each macro-area of knowledge (Section 3.2). Finally, in Section 4, we round off the article by discussing the results and giving some concluding remarks.

2. Network construction and analysis

In this section, we introduce the way of defining the researchers collaboration networks, the TCM algorithm and the metrics used to compare these community partitions with the native (departmental) partition.

2.1 *Network formulation*

The data used to construct the collaboration networks come from the University of Zaragoza published articles and researchers affiliation database, covering a period of time ranging from January 2002 to January 2021, and have been processed by *Kampal Data Solutions* [46] (a spin-off of the University of Zaragoza). We have information of 3844 researchers (we only take into account those researchers who as of January 2021 are still affiliated to the University of Zaragoza) and 111706 published articles. The researchers of the University of Zaragoza are divided in five macro-areas: Sciences, Health Science, Engineering and Architecture, Social Sciences and Human Studies. The way of researching, publishing and rating the articles is significantly different in natural sciences, social sciences and humanities. For this reason, we have centred this work in the study of the three collaboration networks corresponding to the macro-areas of Sciences, Health Science, and Engineering and Architecture. The information included in these data for each researcher is:

- ID number
- Affiliation to macro-area
- Affiliation to department

and for each published paper:

- ID number
- Year of publication
- JCR impact factor
- ID number of authors (researchers).

Using these data, we construct the three undirected and weighted networks, based on the collaboration among researchers through co-authorship in the published articles. Thus, each node in a network corresponds to a researcher affiliated to the corresponding macro-areas, who has published at least one article in a journal whose impact factor appears in the JCR, co-authored with another researcher of the same macro-area. Two nodes are connected by a link if the corresponding researchers have published at least one co-authored article in a journal whose impact factor appears in the JCR, in the case the article has more than two authors, this is applied to each pair of authors. The weight of a link is computed in the following way:

$$w_{ij} = \sum_{m=1}^{M_{ij}} \frac{\mathrm{IF}_{ij}^{m}}{N_{ij}^{m} - 1}\,, \tag{2.1}$$

where w_{ij} is the weight of the link between nodes i and j, M_{ij} is the number of co-authored articles published by researchers corresponding to nodes i and j, IF_{ij}^{m} is the impact factor of the mth co-authored article, and N_{ij}^{m} the number of authors of the mth co-authored article.

In Table 1, it is shown the main information of each of the three networks constructed using this approach, including the number of departments for the departmental partitions of each macro-area.

TABLE 1 *Properties of the scientific collaboration networks analysed*

Macro-area	Nodes	Links	Nodes GC (%)	Links GC (%)	N. of departments
Science	537	1760	94.6	99	14
Health Science	724	3060	97.7	99.7	11
Engineering and Architecture	628	2055	99.4	99.9	10

For each network, we report the number of nodes, *Nodes,* the number of links; *Links,* the percentage of nodes in the Giant Component (percentage with respect to the total number of nodes in the network); *Nodes GC*(%), the percentage of links in the Giant Component (percentage with respect to the total number of links); *Links GC*(%), and the number of departments in the corresponding macro-area, *N. of departments*.

2.2 *Community detection with TCM*

Here, we first introduce the modularity as a measure of the goodness of a community partition and explain the GN algorithm since our algorithm uses the optimal partitions obtained with the GN as input. Afterward, we present the TCM algorithm.

2.2.1 *Modularity and GN algorithm* The modularity [28] of a network community partition accounts for the fraction of edges that fall within each community, minus the expected fraction in the random-distributed case. For a weighted network $\mathscr{G}$, the modularity of a given division into n communities, $\{c_1, \ldots, c_n\}$, is given by:

$$Q(\mathscr{G}; \{c_1, ..., c_n\}) = \frac{1}{2\mathscr{W}_t} \sum_{i,j} \left[w_{ij} - \frac{s_i s_j}{2m} \right] \delta(c_i, c_j) \ , \tag{2.2}$$

where s_i is the strength of node i, $s_i = \sum_j w_{ij}$ and $\mathscr{W}_t$ is the total weight of the network $2\mathscr{W}_t = \sum_i s_i$, and the random edges distribution null model selected is one in which the probability that there exists an edge between nodes i and j is proportional to the product of the strengths of nodes i and j. Thus, Q values can go from -1, when the fraction of edges (weighted) inside communities is the least compared to the random one, to 1 in the opposite case.

The GN algorithm is an iterative, modularity maximization heuristic algorithm, that is the solution found is an approximate solution. The GN algorithm is based on the elimination of edges with the greatest values of Betweenness Centrality (BC) [29, 47]. In each step, the values of BC for all the edges are computed, using an algorithm from the Python library *Networkx*, proposed by Ulrik Brandes [48]; and the edge with the greatest BC value is eliminated. The process ends when there are no edges in the network. During the edges elimination process, the network is divided into disconnected components, which finally are identified with the communities of the partition. Every time a disconnected component is divided, a new partition, with one more community, is obtained. It is important to remark that the algorithm only provides one partition for a given number of communities, this way, we can identify each partition with its number of communities. In the case of a weighted network [49], the BC value is divided by the weight of the corresponding edge before eliminating the edge with the greatest value, in order to avoid the elimination of the heaviest edges within communities. Just as a clarification, to obtain the shortest paths in the BC computation, we consider the inverse of the weight as the length of each edge.

2.2.2 *Targeted Community Merging* The GN algorithm returns then a set of community partitions, each one identified with the number of communities. The optimal partition is just the partition with the highest value of Q; however, this optimal partition usually has a very high number of communities. This fact makes it very difficult to compare this optimal partition with partitions with a much reduced number of communities, such as the departmental partitions. For this reason, we present a heuristic post-processing algorithm called TCM which takes as input the iterative evolution of the community partitions obtained with the GN algorithm (one partition for each number of communities) and returns, after an iterative process, one (or several) optimal partition(s) with a reduced number of communities without significantly reducing Q. This algorithm is based on the idea of detecting and removing community splittings, by merging the formed communities corresponding to low Q increases.

To implement the TCM algorithm, it is necessary to assign an index to the communities resulting from the GN algorithm evolution and track the changes in modularity for each community split. Therefore, in the GN evolution we start by assigning the index $i = 0$ to the first community (the whole network), and in each split (one old community is divided into two new ones), we call the new community with the higher number of nodes the *parent* community, which is identified with the same index as the old community ($i_{parent} = i_{old}$); and the smaller one is the *child* community, identified with a new index ($i_{child} = \mathscr{C}$, where $\mathscr{C}$ is the number of existing communities at the moment of split). Moreover, we assign the variation of modularity of the split ΔQ_i to the 'birth' of the *child* community i. With this information from the evolution of community partition used as input, the TCM algorithm consists of the following:

1. We fixed a minimum modularity variation $\Delta Q_{\min}$
2. Starting with the newest community (the child of the last division of the GN algorithm evolution), we go back over all communities sequentially and mark the communities as *relevant*, which is not going to be rejoined, or *irrelevant*, which is going to be joined to another community. One community i is considered *relevant* if fulfils one of the following conditions:

   $\Delta Q_i > \Delta Q_{\min}$, the birth of the community implies an important growth of Q. In this case, the parent of the community i is automatically marked as *relevant*, without waiting for its 'turn'.

   The community has a *relevant* child

   and is considered *irrelevant* in any other case. An *irrelevant* community is immediately, at the moment of marking, joined to its parent community, if the latter is not *relevant* (*irrelevant* or not yet classified).
3. After this first filtering, we go once more over all *irrelevant* communities that remain unjoined, starting again with the newest. These remaining *irrelevant* communities are those that could not be joined to its parent because the latter was relevant. We join each one of these *irrelevant* communities to the community of the partition which implies the lowest decrease of Q , no matter if it is *relevant* or not.

See Fig. 1 for a graphical explanation of the algorithm. The parameter $\Delta Q_{\min}$ controls the balance between the number of communities and the reduction in Q. Increasing the value of $\Delta Q_{\min}$ the post-processing returns an optimized partition with fewer communities by reducing Q. This implies that we can obtain a more global landscape of several optimized partitions by scanning $\Delta Q_{\min}$ as it is shown in Fig. 3 (Section 3), this variety brings the possibility of choosing an optimal partition with a balance

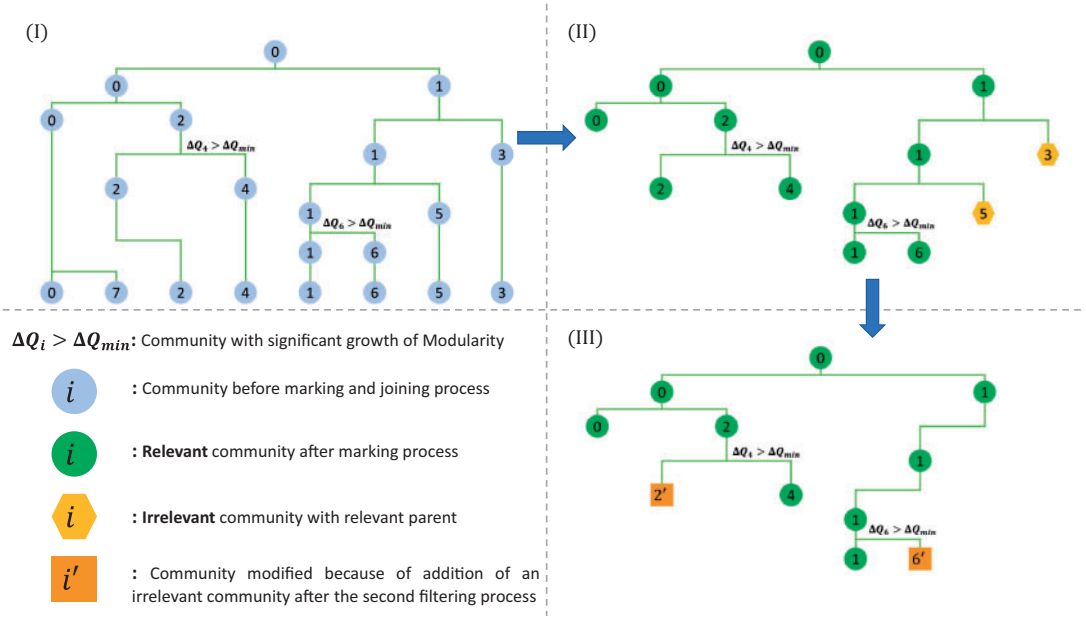

FIG. 1. Graphical explanation of the TCM algorithm using a simple synthetic case. (I) Performance of the usual GN algorithm (from top to the bottom) obtaining an eight community partition by splitting one of the existing communities in each step, in this case, we assume that only the 'birth' of communities four and six implies a relevant growth of Q for the fixed $\Delta Q_{\min}$ ($\Delta Q_i > \Delta Q_{\min}$). (II) First marking and filtering process, community 7 has joined to its parent community because in its turn both were irrelevant, communities 4 and 6 are marked as relevant in their turn, as well as the respective parents (2 and 1), and communities 3 and 5 cannot be immediately joined to their parent, even though they are irrelevant, since in their turns community 1 was already marked as relevant. (III) Second filtering, we check which are the remaining communities to which to join 3 and 5 leading to a minimum decrease in Q, in this case, the best candidate are the 2 for community 5 and 6 for community 3. After the whole post-processing, we end up with a reduced partition with five communities.

between modularity and number of communities which fits a specific problem needs. Moreover, there are two partitions that are interesting in all cases, because of the outperforming of the TCM algorithm:

- The partition obtained by fixing $\Delta Q_{\min} = 0$. All the splittings of the GN algorithm that bring a decrease of Q are eliminated, this way we obtain a partition with the highest possible modularity
- The partition resulting from pushing $\Delta Q_{\min}$ to the limit of obtaining a value of Q just above or equal to the maximum of GN. This way we obtain a partition at least as good as that of the GN algorithm, but with a smaller number of communities.

However, in this work, we do not focus on these two partitions, we perform a more detailed analysis in which we do not only look at modularity. We would like to propose a partition of each macro-area that eventually could be used as a basis for restructuring the current departmental partitions, more based on scientific collaboration structures than simply on branches of knowledge. However, we are interested in partitions that are not too much different from the present ones. For this purpose, we need a way to measure the similarity between different partitions of the same set of objects, in our case we will use some well-known similarity indices.

2.3 *Partition similarity indices*

Once we obtain the optimal community partitions using the TCM algorithm the idea is to compare these partitions with the departmental partition for each network. Similarity indices give us quantitative information about how similar are two partitions. We have used two well-known pair-based similarity indices: Rand and Wallace Index [44, 45]. Besides, we introduce the idea of adjusted for chance correction for the indices, which is applied to the Rand Index (RI). Although there exist other possibilities for the similarity measures, such as those based on information theory [50, 51], we concluded that pair-based measures are more appropriate for this problem.

2.3.1 *Pair-based indices* Given a set of N objects and two different partitions of this set, $\mathscr{U}$ (with U groups) and $\mathscr{V}$ (with V groups), we define $n_{i\cdot}$ as the number of objects that are in group i in partition $\mathscr{U}$, $n_{\cdot j}$ as the number of objects that are in group j in partition $\mathscr{V}$, and n_{ij} as the number of objects that are in group i in partition $\mathscr{U}$ and in group j in partition $\mathscr{V}$. With these definitions, we present the following three pair-based measures:

$$P = \sum_{i=1}^{U} \frac{n_{i\cdot}\,(n_{i\cdot} - 1)}{2}, \tag{2.3}$$

$$Q = \sum_{j=1}^{V} \frac{n_{\cdot j}\,(n_{\cdot j} - 1)}{2}\,, \tag{2.4}$$

$$T = \sum_{i=1}^{U}\sum_{j=1}^{V} \frac{n_{ij}\,(n_{ij} - 1)}{2}, \tag{2.5}$$

where P is the number of pairs of objects that are in the same group in first partition, Q is the number of pairs of objects that are in the same group in second partition, and T is the number of pairs of objects that are in the same group in both partitions. Notice that $\binom{N}{2} = \frac{N(N-1)}{2}$ is the total number of pairs of objects, and we can express the number of pairs of objects that are in different groups in both partitions as $\binom{N}{2} - P - Q + T$.

Given these numbers, the Wallace Index is expressed as follows:

$$W = \sqrt{\frac{T}{P}} \cdot \sqrt{\frac{T}{Q}} = \frac{T}{\sqrt{Q \cdot P}}\,, \tag{2.6}$$

that is, the geometric mean of the fraction $\frac{T}{P}$ of nodes that being in the same group in first partition also are in the same group in the second partition, and the fraction $\frac{T}{Q}$ of nodes that being in the same group in the second partition also are in the same group in the first partition. And the RI is computed as:

$$\mathrm{RI} = \frac{T + (\binom{N}{2} - P - Q + T)}{\binom{N}{2}} = \frac{2T + (\binom{N}{2} - P - Q)}{\binom{N}{2}}\,, \tag{2.7}$$

that is, the sum of pairs of objects that are in the same group in both partitions and the pairs of objects that are in different groups in both partitions, divided by the total number of pairs of objects. The RI assumes that the similarity of two partitions is given by both types of pairs of objects.

In our case, the objects are the researchers (nodes in the networks), the first partition is the community partition, and the second partition is the departmental partition.

2.3.2 *Adjusted for chance correction of indices* Even though these indices are two of the most used in clustering comparison, several studies [52–55] have proven that the performance of the RI is clearly improved by making a *Correction for Chance* of the index.

The idea of the *Correction for Chance* is to eliminate the influence of the randomness grouping of objects on the index. Thus, the Adjusted Rand Index (ARI) is expressed as:

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\,\mathrm{RI}\,]}{\mathrm{RI}_{\max} - E[\,\mathrm{RI}\,]}\,, \tag{2.8}$$

where $E[\,\mathrm{RI}\,]$ is the expected value and $\mathrm{RI}_{\max}$ is the maximum value of RI, both based on the *generalized hypergeometric distribution* null model [52] for randomness which assumes that $\mathscr{U}$ and $\mathscr{V}$ partitions are picked at random, subject to having the original number of groups and objects in each partition. Therefore, we can express these quantities as [52]:

$$E[\,\mathrm{RI}\,] = E\left[\frac{2T + \left(\binom{N}{2} - P - Q\right)}{\binom{N}{2}}\right] = \frac{2\,E[\,T\,] + \left(\binom{N}{2} - P - Q\right)}{\binom{N}{2}}\,, \tag{2.9}$$

where

$$E[\,T\,] = E\left[\sum_{i=1}^{U}\sum_{j=1}^{V}\binom{n_{ij}}{2}\right] = \sum_{i=1}^{U}\sum_{j=1}^{V}\frac{\binom{n_{i\cdot}}{2}\binom{n_{\cdot j}}{2}}{\binom{N}{2}} = \frac{P\,Q}{\binom{N}{2}} \tag{2.10}$$

since the expected number of pairs of objects which belongs to $\mathscr{U}_i$ and $\mathscr{V}_j$ is equal to the product of pairs of objects of $\mathscr{U}_i$ and the pairs of objects of $\mathscr{V}_j$, divided by the total number of pairs of objects, that is $E\left[\binom{n_{ij}}{2}\right] = \frac{\binom{n_{i\cdot}}{2}\binom{n_{\cdot j}}{2}}{\binom{N}{2}}$, and

$$\mathrm{RI}_{\max} = \frac{2\,T_{max} + \left(\binom{N}{2} - P - Q\right)}{\binom{N}{2}}\,, \tag{2.11}$$

while there is no complete consensus in the literature on the expression of $T_{\max}$, being some of the possibilities Q, P, $\min[Q,P]$ or $\frac{P+Q}{2}$, in general cases the preferred choice is $\frac{P+Q}{2}$. Hence, the final expression of ARI, substituting RI from Eq. 2.7, $E[\,\mathrm{RI}\,]$ from Eqs. 2.9 and 2.10, and $\mathrm{RI}_{\max}$ from Eq. 2.11 being $T_{\max} = \frac{P+Q}{2}$; stays as

$$ARI = \frac{T - \frac{P\,Q}{\binom{N}{2}}}{\frac{P+Q}{2} - \frac{P\,Q}{\binom{N}{2}}}\,. \tag{2.12}$$

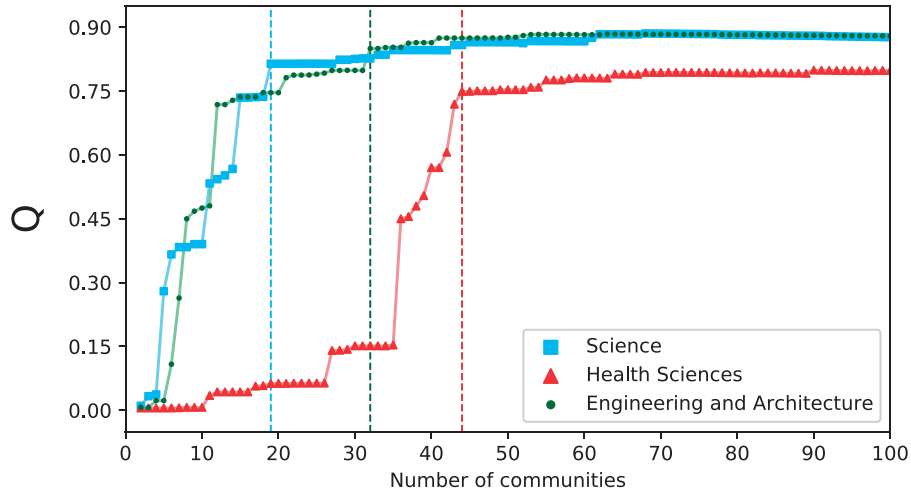

FIG. 2. Performance of the GN algorithm for the detection of communities in each collaboration network. We show the values of modularity for each partition into communities which correspond to one value of the Number of Communities (x-axis). Light blue squares correspond to Science network, red triangles correspond to Health Science and dark green points correspond to Engineering and Architecture. The vertical dashed lines mark the beginning of the steady zone for each macro-area.

This adjustment implies that the ARI value is always strictly greater than -1 and strictly less than 1 and is equal to 0 when the similarity matches what would be expected by chance according to the null mode. For more information about the ARI extreme values and the *generalized hypergeometric distribution* null model see the reference [52–55].

3. Results

We need to obtain the community partitions to be compared to the native (departmental) ones. For this purpose, we applied the TCM algorithm, presented in the section above, to each of the networks of the three macro-areas under study.

This section starts with the results of the performance of the TCM algorithm compared to the usual GN algorithm, then we show the set of community partitions obtained with the new algorithm and the selection of the 'optimal' partition for each macro-area network, based on modularity and similarity criteria, which are going to be qualitatively compared with the departmental partitions.

3.1 *Performance of TCM algorithm*

In Fig. 2, we show the performance of the original GN algorithm for the three macro-areas, representing the modularity of each obtained community partition (uniquely determined by its number of communities). The forward direction of the x-axis corresponds to the evolution of the algorithm, increasing the number of communities by one at each step. We observe a marked growth at the beginning, with some small plateaus; until it reaches a more steady zone (the beginning of these steady zones is marked with vertical dashed lines in the figure) where it takes many steps to achieve the maximum, around the partition of 70 communities in all cases.

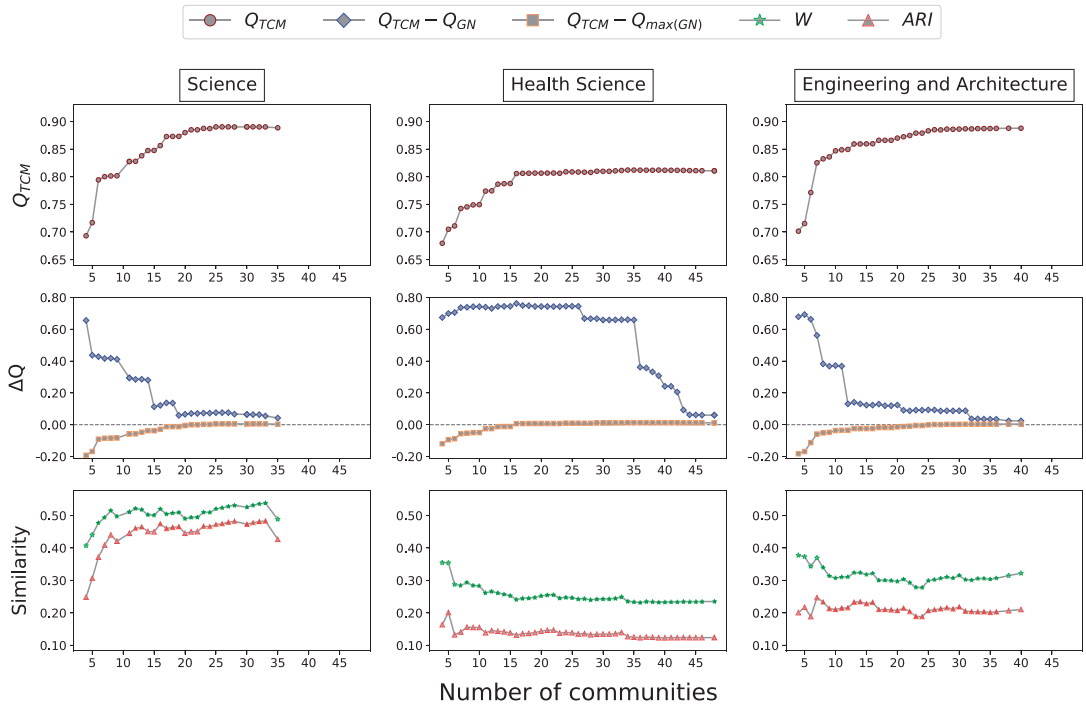

FIG. 3. Performance of TCM algorithm on the three collaboration networks (left: Science, Centre: Health Science and right: Engineering and Architecture). In the upper row, we represent the values of modularity (maroon circles) for each partition into communities, obtained with TCM algorithm, which correspond to one value of the Number of Communities (x-axis). In the centre row, we compare the performance of TCM and GN algorithms by showing the difference between the modularity of the partition obtained with the TCM and the modularity of the partition obtained with the GN for each number of communities (blue diamonds), and by showing the difference of the modularity of the partition obtained with the TCM for each number of communities and the maximum modularity obtained with the GN (sandy-brown squares). In the bottom row, we show the similarity values, Wallace (green stars) and Adjusted Rand Index (red triangles), between community (obtained with TCM) and departmental partitions.

Figure 3 corresponds to the application of the TCM algorithm. Here, the different community partitions (again uniquely determined by the number of communities) correspond to different values of $\Delta Q_{\min}$. For low values of $\Delta Q_{\min}$ (right part of the graphs) the number of communities that have been rejoined is small, implying a small decrease in modularity and a high number of communities. On the contrary, high values of $\Delta Q_{\min}$ imply a greater reduction of communities and modularity (see upper row of the figure). We observe that this algorithm outperforms the usual GN algorithm (see centre row of the figure), we obtain a partition with the same value of modularity as the maximum of GN for a number of communities around 25 for Science, and Engineer and Architecture and just 15 for Health Science; and from that point on all the partitions have more modularity than the GN maximum.

3.2 *Partitions comparison in collaboration networks*

Once checked the performance of the TCM algorithm, we must choose the optimal community partition, for each macro-area network. For this purpose, we will take into account not only the modularity but also the similarity with the departmental partition, as it is mentioned in the previous section. Therefore,

in the bottom row of Fig. 3, we show the similarity between the departmental partition (Adjusted Rand and Wallace indices) ant the output community partitions of the TCM algorithm, for each macro-area.

In the three macro-areas, we see that both similarity indices have similar behaviour, with the difference that the Wallace index always takes slightly higher values. In the case of Engineering and Architecture and Health Sciences, both indices have a peak at the beginning and then slightly decrease to a steady zone. In Science network, on the contrary, at the beginning there is an increase in similarity, although reaching a steady zone as well. What is important is that for the three cases, in the zone of interest (high modularities and close to the maximum value of the usual GN) there are no large variations of similarity. Taking this into account, our criterion for choosing the best partition has been qualitative and based more on the modularity.

For the Health Sciences network, we have chosen the 16 community partition as the optimal one since it has the least number of communities before the modularity is lower than the usual GN and starts a very steep decrease. For the case of Sciences and Engineering and Architecture, we cannot take the smallest partition with a modularity value higher than that of the usual GN, since that is given for a very large number of communities, therefore, we choose qualitatively those partitions that have the least number of communities before a sharp decrease in modularity, which are the 17 communities one for Sciences and that of 13 communities for Engineering and Architecture.

For the qualitative comparison of the optimal community partitions and the departmental partitions, for each network, we make a graphical crossing between both partitions using barplots. In this way, it is possible to obtain information, in a very visual way, about how is the collaboration of researchers within and between the different departments of each macro-area.

In Fig. 4, we represent the comparison of the optimal partition of 17 communities with the 14 departments partition for the Science network. The bars represent the communities, the height is the number of researchers in each community, and the colours represent the departments (in the legend). We notice at first glance that the departments are not very segregated in communities, a department appears at most in three communities and a community is made up of three departments at most. Communities $1, 2, 6, 8$ are those with more mixture, being:

- Community 1 formed by a main part of *Applied mathematics* and parts of *Analytic chemistry* and *Mathematics*
- Community 2 represents the collaboration among *Condensed matter physics* (main part), *Theoretical Physics* and *Earth Science*
- Community 6 is the last of the biggest communities and it is formed by *Organic Chemistry* and *Inorganic Chemistry* almost at equal parts
- Community 8 is the most mixed with parts of *Inorganic Chemistry*, *Mathematics* and *Organic chemistry*,

The rest of the communities are hardly mixed, and each of them can be uniquely identified with a department.

In the comparison of the optimal partition of 16 communities with the 11 departments of the Health Science macro-area, see Fig. 5, we observed a different behaviour from that of the Science macro-area. There are a reduced number of communities which can be uniquely identified with a department, namely: $2, 4, 10, 14$ and 16, but three of these $(4, 10, 14)$ correspond to the same department, *Surgery*. The remaining 11 communities present collaboration between departments, and although in three of

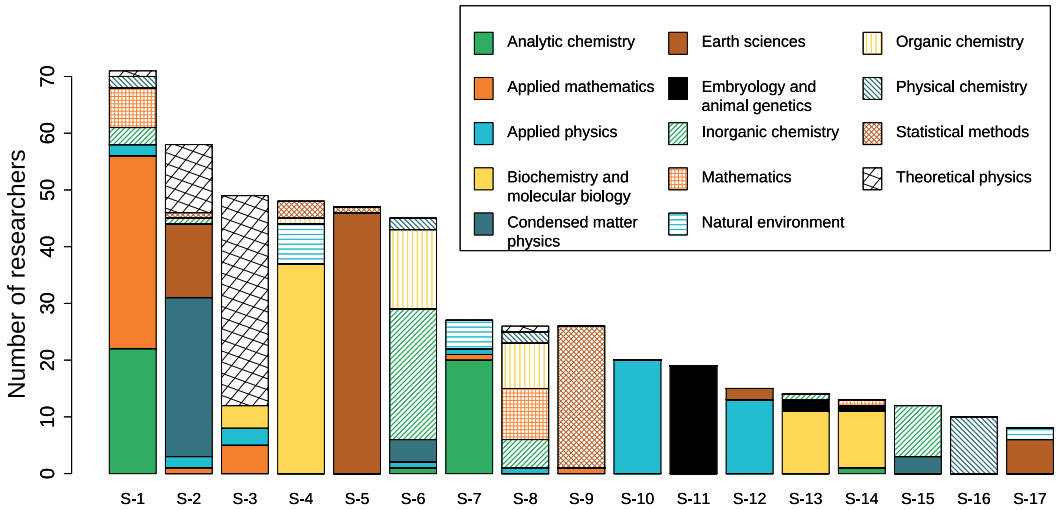

FIG. 4. Barplot comparing the optimal 17 community partition with the departmental partition for the Science macro-area collaboration network. Bars correspond to communities, ordered from left to right in order from highest to lowest number of researchers, and colours represent departments (striped textures used to avoid colour repetition). Y-axis represents the number of researchers per community.

them (9, 12 and 13) the collaboration is only between two departments: *Medicine, psychiatry and dermatology—Surgery*; *Microbiology and public health—Medicine, psychiatry and dermatology*; and *Animal pathology—Animal production* respectively, in the rest there is a great mix of departments, being three the minimum number of departments that appear in them. In addition, it is interesting that there is a pattern of interdepartmental collaboration that is repeated in several communities of different sizes, collaboration among *Surgery*; *Physiatry and nursing*; *Medicine, psychiatry and dermatology*; *Microbiology and public health*; and *Animal pathology* can be observed in communities 1, 6, 7, 8 and 15.

Finally, we show in Fig. 6 the comparison between the optimal partition of 13 communities and the 10 departments partition for the macro-area of Engineering and Architecture. In this case, inter- and intra-departmental collaboration presents a landscape that is midway between the two other macro-areas. There are several communities that can be almost uniquely identified with a department (3, 10, 11, 12, 13) and, as in Science macro-area, all these departments are different. In the case of communities which present departmental collaboration, there is not a great mixture but a little more than in Science, since in some of the biggest communities (1, 2, 4, 6) around four departments are clearly represented. As in Health Science, there is a pattern of collaboration (*Computer science and systems engineering—Electronic engineering and communications*) that is repeated in several communities (1, 2, 4, 5, 7), although in this case, in all these communities, these two are accompanied by different departments.

These differences in behaviour between the three macro-areas in terms of the mixture of inter- and intra-departmental collaboration can be better observed in a more quantitative way with the value of modularity of the partition into departments (Q_{depts}) and with the similarity with departments of the three 'optimal' community partitions $\text{ARI}_{\text{optimal}}$. The modularity of the departmental partition tells us how good would be a network partition in which the communities were precisely the departments, there is obviously a high correlation between these two measures, since a 'good' departmental partition is more

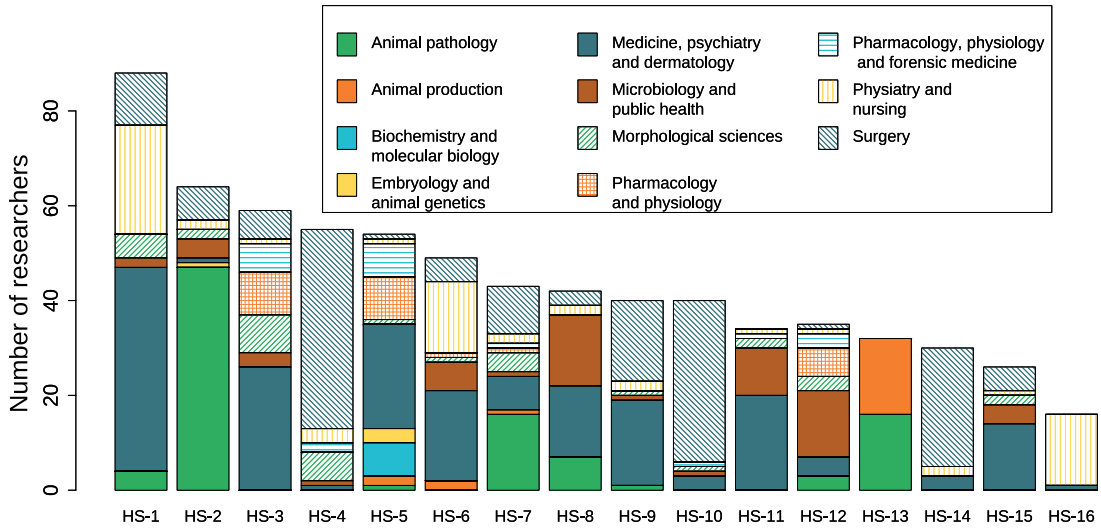

FIG. 5. Barplot comparing the optimal 16 community partition with the departmental partition for the Health Science macro-area collaboration network. Bars correspond to communities, ordered from left to right in order from highest to lowest number of researchers, and colours represent departments (striped textures used to avoid colour repetition). Y-axis represents the number of researchers per community.

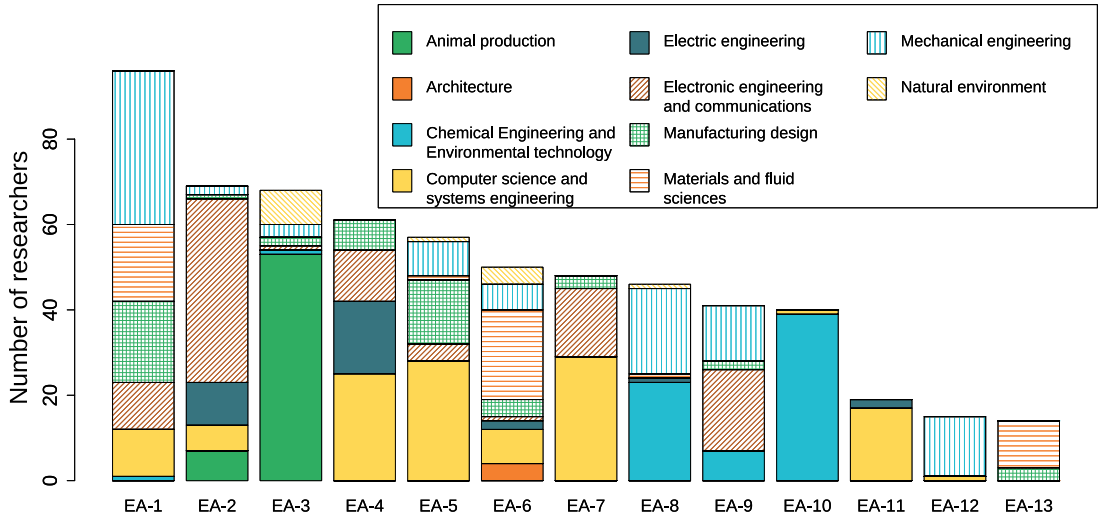

FIG. 6. Barplot comparing the optimal 13 community partition with the departmental partition for the Engineering and Architecture macro-area collaboration network. Bars correspond to communities, ordered from left to right in order from highest to lowest number of researchers, and colours represent departments. Y-axis represents the number of researchers per community.

likely to resemble the optimal community partition found by our algorithm. In Table 2, we show these two measures for the three macro-areas. We can observe, as expected, that the macro-area of Science has the greatest values for both measures, followed by Engineering and Architecture, and the values for Health Science are clearly the lowest.

TABLE 2 *This table shows the values of modularity of departmental partition, and Similarity between optimal community and departmental partitions for the three collaborations networks*

	Q_{depts}	ARI_{optimal}
Science	0.79	0.46
Engineering and Architecture	0.75	0.23
Health Sciences	0.54	0.13

4. Conclusions and future work

The importance of the study of the structural properties of complex networks, where the detection and analysis of community structures stands out, generates the need for continuous research and development of algorithms in this field, especially when applied to the characterization of human-kind collaboration networks.

In this article, we develop a community detection algorithm which improves the performance of the GN algorithm and allows reducing the number of communities without reducing the goodness of the partition. We apply this algorithm to different macro-areas of knowledge at the University of Zaragoza in order to analyse the collaborative structure of the departmental partition of these macro-areas. For this purpose, we make a comparison of each partition in departments with the partition in communities obtained from each collaborative network.

The comparison between the performance of the GN algorithm and the TCM algorithm makes tangible the usefulness of the latter to obtain much smaller partitions (more suitable for comparison with native partitions of networks) with little harm to modularity. Not only that, even slightly relaxing the condition of a reduced number of communities, we can obtain higher modularity values than in the usual algorithm.

The collaborative structure of the departmental partitions presents different features depending on the macro-area. These differences are manifested mainly in the mixture of departments for each macro-area, going from the less mixed partition for Science to the more mixed for Health Science. This could indicate that some partitions into departments were designed taking into account some scientific collaboration criterion while the other might be more based simply on a criterion of branches of knowledge. But, more probably, it could also be a consequence of the fact that in Science it is possible to investigate with the focus on a specific field while in Health Science it is necessary, in general, the participation of several disciplines.

Besides, the emergence of repeating patterns of collaboration in each macro-area can be observed. Some of which might be expected due to their simplicity and the natural relationship among the departments forming them, such as: *Organic chemistry* and *Inorganic chemistry* in Science, or *Computer science and systems engineering—Electronic engineering and communications* in Engineering and Architecture. However, others need deeper unravelling as *Surgery*; *Physiatry and nursing*; *Medicine, psychiatry and dermatology*; *Microbiology and public health*; and *Animal pathology* in Health Science, since it is not a simple pattern and the relationship among these departments is not so obvious.

Definitely, our algorithm has proven to be a very useful tool to obtain good community partitions with a small number of communities, comparable to the typical sizes of native partitions of collaborative scientific networks among others, and allows us to perform meaningful analysis of this type of structures. Besides, the method presented here could help to take decisions about an eventual

restructuring of researching groups or units within an academic or researching institution. Two particular examples:

- The situation where the institutions must reduce management and administrative costs, and therefore it may be necessary to decrease the number of departments or areas. To accomplish this, the TCM algorithm can be used to obtain a community partition, with the specific number of communities required by the institution, which can then be directly associated with the new departments.
- The scenario where an academic institution intends to create a structure of research institutes from scratch, the number of which is usually smaller than the number of departments, and wants to know to which institute it should assign each researcher in the institution. The TCM could be used to obtain the community partition with the same number of communities than the number of institutes that is intended to create.

However, the algorithm and the way of analysing the departmental structure have some limitations that may provide an opportunity for future work. On one hand, it would be interesting to consider more bindings in the post-processing beyond simply looking for the smallest decrease in modularity, or to impose constraints on the communities that the algorithm joins, based on an *a priori* analysis of the native partition to be analysed. On the other hand, further work would be needed in order to include Humanities and Social Sciences in this kind of analysis. The great variability and the lack of a standardized indexation of the journals of these disciplines make it very difficult to integrate them with the scientific fields in a common basis. The present efforts of some academic publishing companies in order to include human and social disciplines in the same indexation schemes as the scientific ones could help to work on this integration in the future.

Funding

The Spanish Ministerio de Ciencia e Innovación (projects PGC2018-094684-B-C22 and PID2020-113582GB-I00); Departamento de Industria e Innovación del Gobierno de Aragón y Fondo Social Europeo through projects no. E30_20R (COMPHYS group) and E36_20R (FENOL group).

Data and code availability

Data from University of Zaragoza used in this study are available upon proper request to *Kampal Data Solutions*: info@kampal.com. The code is publicly available in github: https://github.com/Francho22/TCM_algorithm.git. The reader will find, apart from a Python script correponding to the TCM algorithm, two Python libraries with some methods required by the TCM algorithm, and two *.txt* files as example of the input for the TCM algorithm.

References

1. Estrada, E. (2011) *The Structure of Complex Networks*. Oxford: Oxford University Press.
2. Latora, V., Nicosia, V. & Russo, G. (2017) *Complex Networks: Principles, Methods and Applications*. Cambridge: Cambridge University Press.
3. Newman, M. E. J. (2018) *Networks*. Oxford: Oxford University Press.
4. Boguna, M., Krioukov, D. & Claffy, K. C. (2009) Navigability of complex networks. *Nat. Phys.*, **5**, 74–80.
5. Colizza, V., Flammini, A., Serrano, M. A. & Vespignani, A. (2006) Detecting rich-club ordering in complex networks. *Nat. Phys.*, **2**, 110–115.

6. Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. (2005) Comparing community structure identification. *J. Stat. Mech.*, **2005**, P09008.

7. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. (2006) K-core organization of complex networks. *Phys. Rev. Lett.*, **96**, 040601.

8. Estrada, E., Hatano, N. & Benzi, M. (2012) The physics of communicability in complex networks. *Phys. Rep.*, **514**, 89–119.

9. Fortunato, S. (2010) Community detection in graphs. *Phys. Rep.*, **486**, 75–174.

10. Fortunato, S. & Hric, D. (2016) Community detection in networks: a user guide. *Phys. Rep.*, **659**, 1–44.

11. Lacasa, L. & Gómez-Gardeñes, J. (2013) Correlation dimension of complex networks. *Phys. Rev. Lett.*, **110**, 168703.

12. Song, C., Havlin, S. & Makse, H. (2005) Self-similarity of complex networks. *Nature*, **433**, 392.

13. Stauffer, D. & Aharony, A. (2018) *Introduction to Percolation Theory*. London: Taylor & Francis.

14. Arenas, A., Diaz-Guilera, A. & Pérez-Vicente, C. J. (2006) Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.*, **96**, 114102.

15. Castellano, C., Fortunato, S. & Loreto, V. (2009) Statistical physics of social dynamics. *Rev. Mod. Phys.*, **81**, 591–646.

16. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. (2008) Critical phenomena in complex networks. *Rev. Mod. Phys.*, **80**, 1275–1335.

17. D'Souza, R. M., Gómez-Gardeñes, J., Nagler, J. & Arenas, A. (2019) Explosive phenomena in complex networks. *Adv. Phys.*, **68**, 123–223.

18. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. (2015) Epidemic processes in complex networks. *Rev. Mod. Phys.*, **87**, 925–979.

19. Perc, M., Gómez-Gardeñes, J., Szolnoki, A., Floría, L.M. & Moreno, Y. (2013) Evolutionary dynamics of group interactions on structured populations: a review. *J. R. Soc. Interface*, **10**, 20120997.

20. Szabó, G. & Fáth, G. (2007) Evolutionary games on graphs. *Phys. Rep.*, **446**, 97–216.

21. Arenas, A., Danon, L., Diaz-Guilera, A., Gleiser, P. M. & Guimera, R. (2004) Community analysis in social networks. *Eur. Phys. J. B*, **38**, 373–380.

22. Guimera, R. & Amaral, L. A. N. (2005) Functional cartography of complex metabolic networks. *Nature*, **433**, 895–900.

23. Guimera, R., Mossa, S., Turtschi, A. & Amaral, L. A. N. (2005) The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. USA*, **102**, 7794–7799.

24. Agarwal, G. & Kempe, D. (2008) Modularity-maximizing graph communities via mathematical programming. *Eur. Phys. J. B*, **66**, 409–418.

25. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **2008**, P10008.

26. Massen, C. P. & Doye, J. P. (2006) Thermodynamics of community structure. arXiv, arXiv:cond-mat/0610077, preprint: not peer reviewed.

27. Newman, M. E. J. (2004) Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, **69**, 066133.

28. Newman, M. E. J. (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, **103**, 8577–8582.

29. Newman, M. E. J. & Girvan, M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 026113.

30. Abbe, E. (2017) Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.*, **18**, 6446–6531.

31. Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983) Stochastic blockmodels: first steps. *Soc. Netw.*, **5**, 109–137.

32. Peixoto, T. P. (2014) Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X*, **4**, 011047.

33. Vallès-Català, T., Massucci, F. A., Guimerà, R. & Sales-Pardo, M. (2016) Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys. Rev. X*, **6**, 011036.

34. Guimera, R., Uzzi, B., Spiro, J. & Amaral, L. A. N. (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science*, **308**, 697–702.

35. Migliano, A. B., Battiston, F., Viguier, S., Page, A. E., Dyble, M., Schlaepfer, R., Smith, D., Astete, L., Ngales, M., Gomez-Gardenes, J. et al. (2020) Hunter-gatherer multilevel sociality accelerates cumulative cultural evolution. *Sci. Adv.*, **6**, eaax5913.

36. Ramasco, J. J., Dorogovtsev, S. N. & Pastor-Satorras, R. (2004) Self-organization of collaboration networks. *Phys. Rev. E*, **70**, 036106.

37. Bauzá, F., Ruiz-Manzanares, G., Pérez-Sienes, L., Tarancón, A., Íñiguez, D. & Gómez-Gardeñes, J. (2020) Analyzing the potential impact of BREXIT on the European research collaboration network. *Chaos*, **30**, 063145.

38. Coccia, M. & Wang, L. (2016) Evolution and convergence of the patterns of international scientific collaboration. *Proc. Natl. Acad. Sci. USA*, **113**, 2057–2061.

39. Newman, M. E. J. (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*, **64**, 016131.

40. Newman, M. E. J. (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, **64**, 016132.

41. Newman, M. E. J. (2001) The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, **98**, 404–409.

42. Newman, M. E. J. (2004) Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. USA*, **101**, 5200–5205.

43. Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D. & Barabási, A.-L. (2018) Science of science. *Science*, **359**, eaao0185.

44. Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

45. Wallace, D. L. (1983) A method for comparing two hierarchical clusterings: comment. *J. Am. Stat. Assoc.*, **78**, 569–576.

46. Alvarez, R., Cahué, E., Clemente-Gallardo, J., Ferrer, A., Iñiguez, D., Mellado, X., Rivero, A., Ruiz, G., Sanz, F., Serrano, E. et al. (2015) Analysis of academic productivity based on complex networks. *Scientometrics*, **104**, 651–672.

47. Freeman, L. C. et al. (2002) Centrality in social networks: conceptual clarification. *Social Network: Critical Concepts in Sociology* (J. Scott ed.). vol. 1. London: Routledge, pp. 238–263.

48. Brandes, U. (2001) A faster algorithm for betweenness centrality. *J. Math. Sociol.*, **25**, 163–177.

49. Newman, M. E. J. (2004) Analysis of weighted networks. *Phys. Rev. E*, **70**, 056131.

50. Strehl, A. & Ghosh, J. (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, **3**, 583–617.

51. Vinh, N. X., Epps, J. & Bailey, J. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, **11**, 2837–2854.

52. Hubert, L. & Arabie, P. (1985) Comparing partitions. *J. Class.*, **2**, 193–218.

53. Santos, J. M. & Embrechts, M. (2009) On the use of the Adjusted Rand Index as a metric for evaluating supervised classification. *Artificial Neural Networks ICANN 2009* (C. Alippi, M. Polycarpou, C. Panayiotou & G. Ellinas eds.). Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 175–184.

54. Steinley, D. (2004) Properties of the Hubert-Arable Adjusted Rand Index. *Psychol. Methods*, **9**, 386–396.

55. Steinley, D., Brusco, M. J. & Hubert, L. (2016) The variance of the adjusted Rand Index. *Psychol. Methods*, **21**, 261–272.