The object of this study is complex networks whose model is undirected weighted ordinary (without loops and multiple edges) graphs. The task to detect communities, that is, partition the set of network nodes into communities, has been considered. It is assumed that such communities should be non-overlapped. At present, there are many approaches to solving this task and, accordingly, many methods that implement it. Methods based on the maximization of the network modularity function have been considered. A modified modularity criterion (function) has been proposed. The value of this criterion explicitly depends on the number of nodes in the communities. The partition of network nodes into communities with maximization by such a criterion is significantly more prone to the detection of small communities, or even singleton-node communities. This property is a key characteristic of the proposed method and is useful if the network being analyzed really has small communities. In addition, the proposed modularity criterion is normalized with respect to the current number of communities. This makes it possible to compare the modularity of network partitions into different numbers of communities. This, in turn, makes it possible to estimate the number of communities that are formed, in cases when this number is not known a priori. A method for partitioning network nodes into communities based on the criterion of maximum modularity has been devised. The corresponding algorithm is suboptimal, belongs to the class of greedy algorithms, and has a low computational complexity – linear with respect to the number of network nodes. As a result, it is fast, so it can be used for network partitioning. The method devised for detecting network communities was tested on classic datasets, which confirmed the effectiveness of the proposed approach

Keywords: network modularity, node communities, network partitioning, assortativeness, problems of high dimensionality

UDC 004.03
DOI: 10.15587/1729-4061.2024.318452

NETWORK COMMUNITY DETECTION USING MODIFIED MODULARITY CRITERION

Vadim Shergin
PhD, Associate Professor*
Sergiy Grinyov*
Larysa Chala
PhD, Associate Professor*
Serhii Udovenko
Corresponding author
Doctor of Technical Scienses, Professor
Department of Informatics and Computer Engineering
Simon Kuznets Kharkiv National University of Economics
Nauky ave., 9a, Kharkiv, Ukraine, 61166
E-mail: serhiy.udovenko@hneu.net
*Department of Artificial Intelligence
Kharkiv National University of Radio Electronics
Nauky ave., 14, Kharkiv, Ukraine, 61166

Received 11.09.2024
Received in revised form 01.11.2024
Accepted 11.12.2024
Published 30.12.2024

How to Cite: *Shergin, V., Grinyov, S., Chala, L., Udovenko, S. (2024). Network community detection using modified modularity criterion. Eastern-European Journal of Enterprise Technologies, 6 (4 (132)), 6–13. https://doi.org/10.15587/1729-4061.2024.318452*

1. Introduction

It is well known that real-world networks are not homogeneous. Network nodes are structured into more or less well-defined communities. For example, communities in social networks [1, 2], which can be based on the commonality of location, language, interest domain, cultural aspects [3]. Communities in citation networks reflect the industry specialization of scholars. An important example is also communities in biological networks [4], which reflect the functionality of elements, as well as biological-ecological communities [5], etc. A common feature of such communities is that the nodes of the network are more closely connected to the nodes of their own community than to the nodes of other communities. In other words, the density of connections of any node with nodes belonging to the same community is higher than the average for that node.

Grouping detection is one of the most important tasks of complex network analysis. Thus, dividing the network into communities makes it possible to scale the network by matching the nodes of the new network to the communities of the original network and thus proceed to consider the network on a larger scale. At the same time, the structure of connections between network elements is specified and clarified, and the scale of communities is revealed. In addition, partitioning the network into communities makes it possible to identify atypical nodes and connections that are critically important from the point of view of the integrity of the network, the dissemination of information in it, etc. and. Thus, devising methods for detecting communities in networks is an urgent practical task.

At the same time, despite the high practical significance of the problem under consideration, there are currently no universal, scientifically based methods for solving it. First of all, this is due to the applied, engineering nature of the problem itself, which usually does not involve a single formalized statement of the problem, as well as the lack of well-founded quality criteria for partitioning the network into communities. The second reason is the wide variety of networks themselves. Thus, the use of some formalized criterion and the development of appropriate methods and algorithms for network partitioning can be scientifically based, recognized,

and used by application specialists. However, over time, these methods and algorithms are "refuted" by a counterexample, i.e., the presence of such a network for which the proposed division will be far from optimal. In addition, an important property of the considered problem is its high computational complexity, which makes it impossible to solve it by the method of complete enumeration even for networks of low dimensionalities. For this reason, some methods, despite a strong mathematical justification, are not used in practice. Conversely, there are, and are successfully applied, semi-empirical methods and approaches that do not have a strict justification but show effectiveness in solving practical tasks. One can say that the inconsistency of the statement and the intractability of the general problem of optimal partitioning of network nodes into communities are the reason and justification for the relevance of scientific research aimed at solving the problems of community detection in complex networks.

2. Literature review and problem statement

The task of partitioning a network into communities is very general. Accordingly, there are many different nuances in the statement of this problem, which reflect the expected properties of the object (that is, the features of the connections in the network and the requirements for the quality of its partitioning). Thus, connections in the network can be directed or undirected; have an arbitrary real weight or, for example, only integral or only binary (0/1); multiple connections and loops may or may not be allowed.

A characteristic classification requirement for partition properties is the admissibility or inadmissibility of intersections of selected communities. Thus, a person, as a node of a network of social ties, simultaneously joins various communities formed by territorial, linguistic, professional, or other characteristics. On the other hand, the intersection of communities may reflect the fact of unclear classification. The simplest for analysis (but also the most practically important) is the case of dividing network nodes into pairwise non-intersecting sets, which is the main subject of further analysis.

For the considered problem, there are three main families of solution methods: based on the maximum likelihood method (MLE), based on centrality, and based on modularity maximization. Methods based on maximum likelihood have a strict statistical justification: they boil down to the maximization of the Kullback-Leibler measure [6] (differences in the distribution between the partitioning of nodes by communities being formed) and some null model describing the random partitioning. An important advantage of such methods is a strict justification of both the methods themselves and the properties of the null model [7], against which the partitioning is carried out. The most important drawback of methods based on maximum likelihood is the rather high computational complexity of the corresponding algorithms [8], which limits their practical applicability. It is important that in all cases the problem of identifying groups of network nodes has a high computational complexity, so its optimal solution for large networks is impossible. This creates the need to use various suboptimal, usually greedy, algorithms [10]. In addition, the variability of the mathematical description of null models is not high [7], and the influence of the used model on the properties of the resulting network division options is not always obvious.

The Girvan-Newman method [4] implements network partitioning by step-by-step removal of edges that have the greatest influence on centrality (in the sense of mediation). This method has a high computational complexity ($O(m^2n)$, where m is the number of edges and n is the number of nodes in the network) and is therefore not applicable to most real-world networks.

In the applied aspect, the most common methods for dividing network nodes are methods based on modularity maximization [9], in particular, the Louvain algorithm [10]. Their main advantage is the relative simplicity of implementation and, accordingly, high productivity. Network modularity is an indicator of the tendency of nodes to group into clusters, also called communities. Typically, networks are heterogeneous, so nodes in the same community are strongly connected to each other, while connections between nodes from different communities are rare. In this case, the modularity indicator is high, otherwise it is low.

The determination of network modularity is based on the use of a random graph as a null model, preserving the degrees of the nodes (that is, the weights of the connections) of the analyzed network. The problem is that the null model (random graph) used in the classical approach does not assume the presence of communities. Moreover, the modularity of communities does not directly depend on the number of nodes in these communities. In this regard, it is proposed to modify the definition of modularity of individual network communities and the network as a whole. The proposed changes will reflect the influence of the number of nodes in the communities, accordingly, maximizing the modified modularity will make it possible to detect small communities if they are present in the network.

The modularity of the network $Q(G)$ is defined as the sum of the modularities of all communities of this network. Thus, the network modularity index is additive with respect to communities and is quite simply calculated. This leads to the widespread use of the modularity index to assess the quality of the division of community network nodes.

At the same time, solving the inverse problem (detecting communities and dividing nodes into communities) by maximizing network modularity is associated with known problems. The first is the impossibility of maximizing $Q(G)$ with respect to C_u by a complete search due to high (subfactorial) complexity, which reaches $O((n/\log n)^n)$. The general problem of high-dimensional experiment planning is considered in [11]. Suboptimal greedy algorithms are usually used to partition nodes into communities by maximizing modularity, the most known of which is the Louvain algorithm [8, 10].

The second problem is the limited resolution [12] of grouping search algorithms based on modularity. This means that they tend to form communities that are close in size and therefore poorly distinguish communities consisting of few nodes, whereas in real networks such may exist. To overcome this shortcoming, an additional parameter is introduced into the community modularity definition formula – the resolution factor $\gamma>0$ (by default 1). If $\gamma<1$, then the algorithm prefers larger communities, otherwise – smaller ones. It is important to note that the value $\gamma\neq1$ violates the statistical meaning of modularity as a deviation of the actual number of connections within the community from the expected one. Moreover, varying the value of the resolution factor does not always make it possible to distinguish small communities.

In addition, in traditional methods of partitioning nodes by maximizing the modularity criterion (for example, the Louvain algorithm [10]), due attention is not paid to the validity of the choice of the value of the number of

communities (K). The simplest approach is that communities stop merging when the modularity of the network stops growing. This approach to choosing the number of communities is not statistically justified.

Thus, the methods for detecting communities in the network, based on the maximization of modularity, although they are the most promising, have serious drawbacks: the presence of statistically unjustified tuning parameters, the unreasonableness of the number of communities, and the reluctance to select small communities.

3. The aim and objectives of the study

The purpose of our study is to devise a method for partitioning network nodes into communities based on a modified modularity criterion. This will make it possible to use it for large networks for their further scaling (zooming), analysis of community structure, anomalous nodes, and other tasks.

The set goal can be achieved by solving the following problems:

– to devise a modified modularity criterion;

– to develop an algorithm for partitioning network nodes into communities based on the maximization of the modified modularity criterion, evaluate the computational complexity of the proposed algorithm;

– to perform an experimental test of the performance of the proposed partitioning method on known test datasets of networks and analyze the quality of the partitioning.

4. The study materials and methods

The object of our research is complex networks whose mathematical model is an undirected weighted graph $G(V,E)$.

The task is to partition a set of network nodes into communities. It is assumed that these communities do not intersect in pairs, so the partition into communities C_u, $u=1,...,K$ is called admissible if these communities form a complete group: $Nodes=\bigcup_{u=1}^{K} C_u$, $\forall u\neq v: C_u\cap C_v=\varnothing$. The number of communities to be searched for (K) is generally assumed to be unknown.

The solution to the given problem is based on the maximization of the objective function $Q(G,C_1,...,C_k)$, which is termed modularity.

This paper provides a definition of the proposed criterion of modularity, different from the traditional criterion. The value of this criterion depends on the properties of the network and the current division of this network into communities. An analysis of this dependence is carried out, which allows us to draw preliminary conclusions about the properties and advantages of the modified modularity criterion.

Since the number of partitions of a set of n elements (known as the Bell number) grows very quickly (subfactorially) with increasing n, the search for the maximum of the objective function (modularity) by the method of direct selection is impossible in the general case. Therefore, two suboptimal methods based on the use of greedy algorithms are proposed to solve the problem of partitioning network nodes. This makes it possible to preserve the main property of network partitioning methods based on modularity – low computational complexity. The Python programming language in the Spyder environment was used for programmatic implementation of the proposed methods.

Experimental verification of the performance of the proposed partitioning methods is carried out using test datasets of networks: Zachary karate club [4] and Newman's Polbooks [13], which are well-known datasets used for testing methods, algorithms, and programs for partitioning networks into communities.

5. Results of research on identifying communities in networks using a modified modularity criterion

5. 1. Modified modularity criterion

In a random undirected graph $G(V,E)$ of order $N=order(G)$ for nodes with weights k_i, $i=1,...,N$, the expected weight of an edge between nodes i,j is equal to $k_ik_j/(2m)$, where $m=\frac{1}{2}\sum_{i=1}^{N}k_i$ is the total weight of network edges.

It should be noted that for unweighted networks, the weights of vertices are equal to their powers $k_i=deg(i)$, the weight of the entire network is equal to the number of edges $m=size(G)$, and the expected weight of an edge between nodes i,j is equal to the probability of the presence of this edge.

For each pair of nodes i, j, it is possible to calculate the deviation of the actual weight of the connection between them (that is, the element of the adjacency matrix $A_{i,j}$ of the weighted graph) from the expected weight of the connection:

$$\Delta_{i,j}=A_{i,j}-\frac{k_ik_j}{2m}. \quad (1)$$

Community modularity (C_u) is equal to the sum of values (1) taken from all nodes included in this community:

$$q_u=\frac{1}{2m}\sum_{i,j\in C_u}\left(A_{i,j}-\frac{k_ik_j}{2m}\right)=\frac{1}{2m}\left(L_u^{in}-\frac{\left(L_u^{tot}\right)^2}{2m}\right), \quad (2)$$

where L_u^{in} is the weight of all links in the community u (for an unweighted graph), which is equal to twice the number of edges, both ends of which belong to the community; $L_u^{tot}=\sum_{i\in C_u}k_i$ is the total weight of all nodes included in the community C_u.

To increase the resolution [12] of algorithms for finding groups based on modularity, an additional parameter is introduced into the community modularity definition formula (2) – the resolution coefficient $\gamma>0$ (by default $\gamma=1$):

$$q_u=\frac{1}{2m}\sum_{i,j\in C_u}\left(A_{i,j}-\gamma\frac{k_ik_j}{2m}\right)=\frac{1}{2m}\left(L_u^{in}-\gamma\frac{\left(L_u^{tot}\right)^2}{2m}\right). \quad (3)$$

If $\gamma<1$, then the algorithm prefers larger communities, otherwise – smaller ones.

Then the formula for calculating network modularity takes the following form:

$$Q(G)=\sum_{u=1}^{K}q_u=\frac{1}{2m}\sum_{u=1}^{K}\left(L_u^{in}-\gamma\frac{\left(L_u^{tot}\right)^2}{2m}\right). \quad (4)$$

It is important to note that the value $\gamma\neq1$ violates the statistical meaning of modularity as a deviation of the actual number of connections within the community from the expected one. Moreover, varying the value of the resolution

factor does not always make it possible to distinguish small communities.

Suppose that the network is divided into K non-intersecting communities C_u with n_u nodes in each of them. The numerical values of K and n_u are assumed to be unknown. Let the total number of network nodes equal $N=\sum_{u=1}^{K} n_u$. According to the null model, any node $i\in C_u$ is connected to other network nodes with equal probability. Then the probability of linking node i with node j from the same community C_u is equal to:

$$p_u=\Pr\left(j\in C_u \mid i\in C_u\right)=\frac{n_u-1}{N-1}. \quad (5)$$

Therefore, the expected weight of connections between node i and other nodes of the class C_u is $\mathrm{E}\left\{L_u^{in}(i)\right\}=p_u k_i$. A natural measure of the modularity of a node can be the difference between the actual weight of the connections of the current node i with the nodes of the class C_u and the expected value:

$$\Delta L_u^{in}(i)=L_u^{in}(i)-\mathrm{E}\left\{L_u^{in}(i)\right\}=L_u^{in}(i)-p_u k_i. \quad (6)$$

It is worth noting that a model similar to (6) was used in [14] to estimate network assortativeness. In addition, the task to measure assortativeness [15–17] is closely related to the problem of community detection [1, 18].

As in the case of the traditional definition, the modularity of the entire community is equal to the normalized sum of the modularities of all nodes in that community:

$$\mu_u=\frac{1}{2m}\sum_{i\in C_u}\left(L_u^{in}(i)-\mathrm{E}\left\{L_u^{in}(i)\right\}\right)=\frac{1}{2m}\left(L_u^{in}-p_u L_u^{tot}\right). \quad (7)$$

To estimate the modularity of the entire network, we sum up the local modularities (7) over all the communities of the network:

$$\mu(G)=\sum_{u=1}^{K}\mu_u=\frac{1}{2m}\sum_{u=1}^{K}\left(L_u^{in}-p_u L_u^{tot}\right). \quad (8)$$

It is easy to see that the obtained modularity coefficient (7) is defined for any nonempty communities. The lower limit of the modularity of the network (8) corresponds to the case $\forall u\colon L_u^{in}=0$ and cannot be less than –1, while the upper limit, achievable for an ideal partition $\left(\forall u\colon L_u^{in}=L_u^{tot}\right)$, does not exceed +1. Therefore, $-1\le\mu(G)\le+1$.

Comparing the proposed definition of modularity (7), (8) with the traditional definition (3), (4), we can conclude that the only (but very significant) difference is due to the method for estimating the expected value of the weight of all links in the u-th community $\mathrm{E}\{L_u^{in}\}$. According to the traditional approach it is equal to $\gamma(L_u^{tot})^2/(2m)$, while in our method this estimate is equal to $p_u L_u^{tot}$. Given the definition p_u (5), it can be concluded that these estimates coincide when using the resolution parameter γ, which is equal to:

$$\gamma_u=p_u\frac{2m}{L_u^{tot}}=\frac{2m}{N-1}\Big/\frac{L_u^{tot}}{n_u-1}\approx\frac{\bar{k}}{\bar{k}_u}, \quad (9)$$

where $\bar{k}$, $\bar{k}_u$ are the average weights of nodes in the entire network and in community u, respectively.

Thus, the proposed method for calculating the modularity of the network, on the one hand, is statistically justified through the probabilistic model (5), and on the other hand, it can be considered as a variant of the conventional method using individual settings of the resolution parameter (9). In both versions, conventional (4) and proposed (8), network modularity is the sum of deviations of the actual number of connections in communities L_u^{in} from the expected one. According to both null models for random distributions of random networks, L_u^{in} is a random variable with an expected value $\gamma(L_u^{tot})^2/(2m)$ and $p_u L_u^{tot}$, accordingly, a finite variance. It follows that the variance of the sum of these deviations (that is, the variance of the value of the modularity of the entire network) is also finite and grows in proportion to the number of terms, that is, communities (K). In addition, as will be demonstrated below, the use of the modularity coefficient in the conventional (unaveraged) form (8) leads to the premature termination of greedy algorithms with the formation of communities of single nodes. Therefore, it is proposed to normalize the modularity of the network (8) by dividing its value by $\sqrt{K}$:

$$\bar{\mu}(G)=\frac{1}{\sqrt{K}}\sum_{u=1}^{K}\mu_u=\frac{1}{2m\cdot\sqrt{K}}\sum_{u=1}^{K}\left(L_u^{in}-p_u L_u^{tot}\right). \quad (10)$$

The proposed normalization makes it possible to compare the modularity of network partitioning for different numbers of communities.

5. 2. Algorithms for partitioning network nodes into communities based on the maximization of the modified modularity criterion

Detecting communities by maximizing modularity (8) or (10) for large values of k and C_i using exhaustive sorting is not possible due to the high complexity of the computations required. Two variants of greedy algorithms were proposed: "moderately greedy" and "very greedy". Their pseudocodes are shown in Fig. 1, 2, respectively. In both algorithms, each node is initially considered as a separate community, that is, the initial modularity values are zero: $\mu(G)=\bar{\mu}(G)=0$. In both algorithms, the findbestcomm4U(u,Comm) procedure is used to select the community C_v to which the current node u is added. In this procedure, all nodes v that are adjacent to u are sorted, and such a community $C_v\in v$ is selected, the joining of u to which maximizes the overall modularity (8) of the network. The current value of the total modularity of the network is denoted as gainU.

```
algorithm medium_greedy(G):
iterate for u in Nodes:
     comm[u] = C[u] = set(u)
end_of_iterate
mod_prev = -inf
mod_curr = 0
while mod_curr > mod_prev:
     random_permutation(comm)
     iterate for u in comm:
          gainU, v = findbestcomm4U(u, C)
          Exclude(u, C[u])
          C[v] = union(C[v], C[u])
     end_of_iterate
     mod_prev = mod_curr
     mod_curr = modularity(C)
     comm = C
end_of_while
return C
```

Fig. 1. Pseudocode of the "moderately greedy" algorithm

In the moderate-greedy algorithm, in a loop through the primary communities $u=1,\ldots,k_{iter}$, immediately after the

current community $u \in Cu$ is found (by calling findbestcomm4U(u,Comm)) such a community Cv, joining u to which maximizes gainU, the operations of joining u to Cv ($Cv=Cv\cup\{u\}$) and removing u from Cu($Cu=Cu\setminus\{u\}$) are performed. At the same time, the community Cu may or may not become empty. After the loop completes, the list of communities is updated, and the next iteration of the outer loop is executed. Therefore, this option essentially coincides with the Louvain algorithm [10] with the replacement of the conventional modularity criterion (4) with criterion (8) or (10).

According to a very greedy algorithm (Fig. 2), a pair of communities Cu, Cv is selected in the community cycle so that their merger leads to the maximum possible increase in criterion (8) or (10). After the loop completes, Cv is updated as $Cv=Cv\cup Cu$, and Cu is removed from the community list. Thus, after each iteration, the number of communities is reduced by one. This process (for the outer loop) continues as long as there are possible mergers (that is, as long as the modularity of the network increases).

```
algorithm very_greedy(G)
iterate for u in Nodes:
     comm[u] = C[u] = set(u)
end_of_iterate
mod_prev = -inf
mod_curr = 0
while mod_curr > mod_prev:
     ubest = u[1]
     gainbest = -inf
     iterate for u in comm:
          gainU, v = findbestcomm4U(u, C)
          if gainU > gainbest:
               gainbest = gainU
               ubest = u
               vbest = v
     end_of_iterate
     Exclude(ubest, C[ubest])
     C[vbest] = union(C[vbest], C[ubest])
     mod_prev = mod_curr
     mod_curr = modularity(C)
     comm = C
end_of_while
return C
```

Fig. 2. Pseudocode of the "very greedy" algorithm

As can be seen from the pseudocodes shown in Fig. 1, 2, the more complex and most deeply nested part of both algorithms is the findbestcomm4U function call. This function iterates through communities that have common edges with the nodes of the current community, so its computational complexity is $O(k)$, where $k \le d\_\max, k \le k_{iter}$, k_{iter} is the current number of communities $d_{\max}$ is the maximum power of the primary community (if we consider it as a node). Thus, the complexity of the moderately greedy algorithm is equal to:

$$T_1 = O(N \cdot k),\ k = \max\{d_{\max}, K\}, \qquad (11)$$

that is, it is linear with respect to the number N of network nodes.

The computational complexity of one iteration of the very greedy algorithm is $O(nk)$, where n is the current number of communities, $n=N, N-1,..., K$. Thus, the complexity of the very greedy algorithm is:

$$T_2 = O\left(N^2 \cdot k\right),\ k = \max\{d_{\max}, K\}. \qquad (12)$$

It is evident that the moderately greedy algorithm is much faster than the very greedy one but its result (both the modularity value and the configuration of communities) depends on the order of traversal of the communities in the loop at $u=1,...,k_{iter}$. On the other hand, a very greedy algorithm, which is independent of the random order of communities, allows tracking the change of network modularity in a step-by-step fashion.

5. 3. Experiments on test networks

In this chapter, the proposed network partitioning method is investigated on test datasets. The first of them is the well-known network dataset of the karate club studied by Zachary [4]. It is an undirected weighted network containing $N=34$ nodes that have a total weight of $2m=462$. This data set, with a known partition into two groups, is widely used in studying the community structure of networks. The nodes of the groups are labeled as "Mr. Hi" and "Officer", and each group consists of 17 nodes.

Maximization of the unaveraged modularity criterion (8) using a very greedy algorithm leads to the division of this network into three large communities (of 14, 10, and 5 nodes) and 5 single nodes ('J', 'S', 'c', 'L' , and 'R'). The achieved value of the criterion $\mu(G)=0.4773$. The dendrogram is shown in Fig. 3. For clarity of visualization of the dendrogram, the digital signs of the nodes (1, ..., 34) are replaced by the letters A, ..., Z, a, ..., h. The values of the height axis on the dendrogram correspond to the values of the modularity criterion.

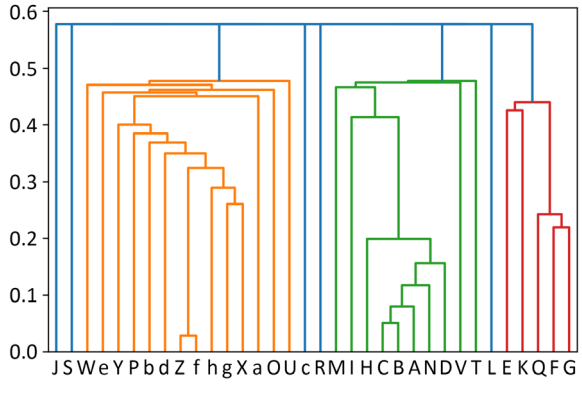

Fig. 3. Dendrogram of karate_club network partitioning by unaveraged modularity criterion (8) using a very greedy partitioning algorithm

It is important to note that the dependence of the unaveraged modularity (8) of the network on the number of communities (Fig. 4) is flat, that is, its maximum is weakly expressed. Thus, the obtained value for the number of communities ($k=8$) is largely determined by the properties of the very greedy algorithm used. When using another partitioning algorithm, one can expect a value of k in the range from 3 to 12.

According to the numerical experiment, the maximization of the same unaveraged modularity criterion (8) using a moderately greedy algorithm leads to the division of this network into two communities (with 16 and 18 nodes). Compared to the partitioning by the very greedy algorithm (Fig. 3), nodes 'J', 'S', 'c' are merged into a community of 14 nodes ("Officer"), nodes 'L' and 'R' joined communities with 10 and 5 nodes ("Mr.Hi"), but node 'I' moved from community "Officer" to "Mr. Hi", which is incorrect. This node 'I' (#9) is known to be problematic for many of the classification procedures tested on this dataset. Thus, the moderately greedy algorithm creates a partition that is close

to ideal, reaching a high value of the criterion $\mu(G)=0.4185$ and is significantly faster than the very greedy variant of the proposed algorithm.

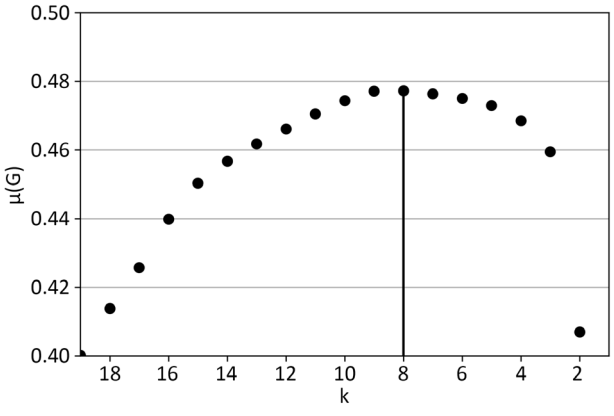

Fig. 4. Dependence of the unaveraged modularity of the karate_club network on the number of detected communities by the very greedy partitioning algorithm

Maximization of the average modularity (10) by a very greedy algorithm leads to the division of this network into two communities (17 nodes each). It is worth noting that this section fully corresponds to the initial node labels ("Mr. Hi" and "Officer"), that is, the resulting partition shown in Fig. 5 is ideal. The achieved value of the criterion $\bar{\mu}(G)=0.2877$, corresponds to the unaveraged value $\mu(G)=0.4069$ and is slightly less than for the non-ideal distribution 16/18.

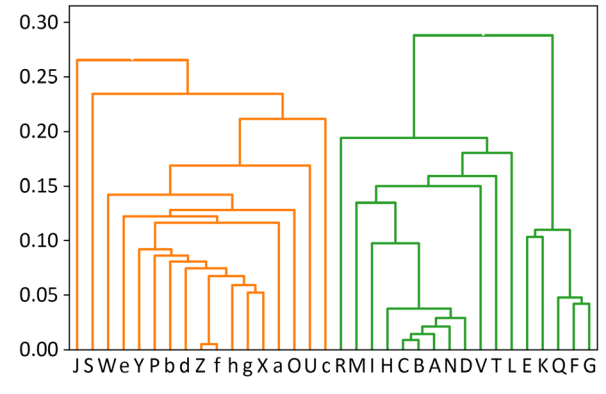

Fig. 5. Dendrogram of karate_club network partitioning according to the average modularity criterion (10) using a very greedy partitioning algorithm

The second test dataset used to investigate the proposed network partitioning method is the popular Newman's Polbooks [13]. The nodes of this set are books on US politics sold by the online bookseller Amazon.com. Nodes (books) are connected by an edge if these books were ordered by the same customer. Nodes are labeled with the letters "*c*", "*l*", and "*n*" (49, 43, and 13 nodes, respectively) to indicate whether they are conservative, liberal, or neutral in their respective political views. Mark Newman assigned these labels separately based on reading the descriptions and reviews of the books posted on Amazon [19]. The network contains 105 nodes and 441 edges, and the maximum degree of a node is 25.

Maximization of the unaveraged modularity criterion (8) by a very greedy algorithm leads to the division of this network into six communities (with 4, 38, 37, 10, 4, and 12 nodes). The achieved value of the criterion $\mu(G)=0.5628$. The dendrogram of the distribution of the network is shown in Fig. 6, and the dependence of the used criterion on the number of detected communities (Fig. 7) is similar to the corresponding dependence for the karate club (Fig. 4).

At the same time, partitioning according to the normalized modularity criterion (10) leads to the division of the network into three communities (Fig. 8) with 38, 41, and 26 nodes, respectively.

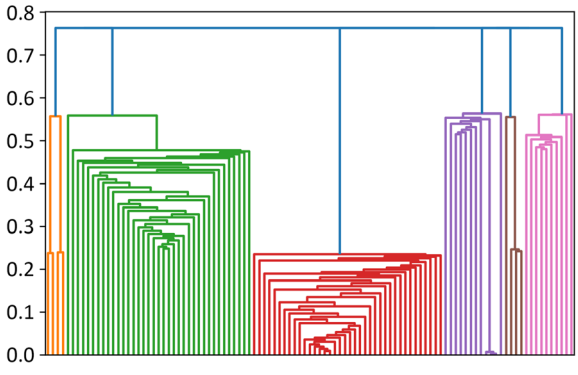

Fig. 6. Dendrogram of partitioning the polbooks network according to the unaveraged modularity criterion (8) using a very greedy partitioning algorithm

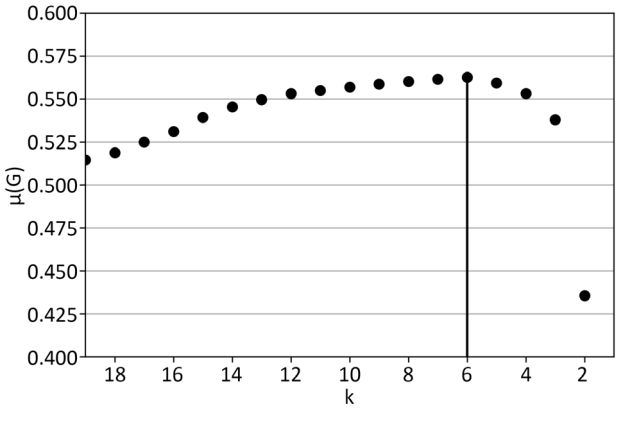

Fig. 7. Dependence of the unaveraged modularity of the polbooks network on the number of detected communities by the very greedy partitioning algorithm

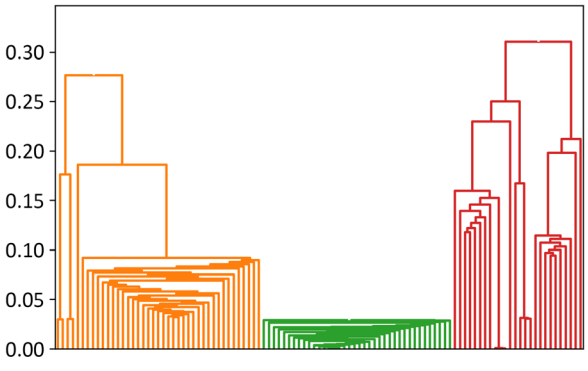

Fig. 8. Dendrogram of partitioning the polbooks network according to the normalized modularity criterion (10) using a very greedy partitioning algorithm

The resulting partitioning cannot be called exact: half of the nodes assigned to community "*n*" actually belong to classes "*c*" and "*l*". At the same time, the number of communities ($K=3$)

corresponds to the actual value. It is worth noting that the original ("real") marking of nodes is subjective. Therefore, the quality of partitioning of the polbooks network by the proposed method can be considered moderate.

6. Discussion of results of research on the detection of communities in networks using a modified modularity criterion

The modified modularity criterion (10) proposed in this study is based on the null model of the network, according to which the probability (5) of a node connecting with another node of its community clearly depends on the number of nodes in this community. The proposed network modularity assessment method differs from the conventional method [9], in which the specified probability depends only on the degrees of connected nodes. As a result, the community detection method based on the maximization of modified modularity has a greater tendency to detect small communities, including single-node ones. This property determines the practical significance of our work. It meets the initial requirements for the devised method and is its advantage over conventional methods [10, 12] with the limitation that the network does contain small communities.

The effectiveness of the proposed method was tested on widely known network datasets. The results of partitioning the network into communities, shown on the dendrograms (Fig. 5, 6, 8), confirm the efficiency of the proposed method: the obtained partitioning is natural and justified, it reflects the objectively existing structure of relationships between the nodes of the analyzed networks.

In addition, the problem of determining the optimal number of communities (K) has been considered. To ensure comparability of modularity values for different K, it has been proposed to normalize the modularity of the network (equal to the sum of modularity of communities) by dividing by the root of K, which was not done in conventional algorithms [8, 10]. The proposed approach partially solves the specified problem (which also has practical significance), but in general it tends to underestimate the number of communities. However, this drawback is common to all community detection methods based on modularity maximization. Therefore, in the practical application of the proposed method, it is recommended to analyze not only the last partitioning of network nodes generated by the algorithm but also previous ones (with larger K values).

It is evident that the task of estimating the optimal number of network communities remains open for further theoretical research.

7. Conclusions

1. A modified modularity criterion has been devised, which clearly depends on the number of nodes in each community. This property increases the tendency to detect small communities (up to single-node communities). In addition, the proposed modularity function is normalized with respect to the number of communities, which makes it more relevant in the case where the number of communities is unknown a priori.

2. Two algorithms for partitioning network nodes into communities using a modified modularity function as an objective function for maximization have been developed. The "very greedy" algorithm has a quadratic complexity with respect to the number of nodes and is used for detailed analysis of the community detection process, as well as for constructing dendrograms. The "moderately greedy" community detection algorithm is much faster (it has a linear complexity with respect to the number of nodes) and is the main variant of the proposed method.

3. The proposed community detection method was tested on widely known network datasets. The results of the experiment as a whole confirm the effectiveness of the proposed method (all karate club nodes and 91 % of polbooks nodes are classified correctly) and the compliance of its properties with the original requirements for computational complexity (it is linear).

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

All data are available, either in numerical or graphical form, in the main text of the manuscript.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

References

1. Newman, M. E. J. (2003). Mixing patterns in networks. Physical Review E, 67 (2). https://doi.org/10.1103/physreve.67.026126
2. Cinelli, M., Peel, L., Iovanella, A., Delvenne, J.-C. (2020). Network constraints on the mixing patterns of binary node metadata. Physical Review E, 102 (6). https://doi.org/10.1103/physreve.102.062310
3. Hamdaqa, M., Tahvildari, L., LaChapelle, N., Campbell, B. (2014). Cultural scene detection using reverse Louvain optimization. Science of Computer Programming, 95, 44–72. https://doi.org/10.1016/j.scico.2014.01.006
4. Girvan, M., Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99 (12), 7821–7826. https://doi.org/10.1073/pnas.122653799
5. Pascual-García, A., Bell, T. (2020). functionInk: An efficient method to detect functional groups in multidimensional networks reveals the hidden structure of ecological communities. Methods in Ecology and Evolution, 11 (7), 804–817. https://doi.org/10.1111/2041-210x.13377

6. Newman, M. E. J. (2004). Detecting community structure in networks. The European Physical Journal B – Condensed Matter, 38 (2), 321–330. https://doi.org/10.1140/epjb/e2004-00124-y
7. Karrer, B., Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. Physical Review E, 83 (1). https://doi.org/10.1103/physreve.83.016107
8. Cohen-Addad, V., Kosowski, A., Mallmann-Trenn, F., Saulpic, D. (2020). On the Power of Louvain in the Stochastic Block Model. Advances in Neural Information Processing Systems (NeurIPS 2020). Vancourer, 4055–4066. Available at: https://hal.science/hal-03140367
9. Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103 (23), 8577–8582. https://doi.org/10.1073/pnas.0601602103
10. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008 (10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008
11. Raskin, L., Sira, O. (2021). Devising methods for planning a multifactorial multilevel experiment with high dimensionality. Eastern-European Journal of Enterprise Technologies, 5 (4 (113)), 64–72. https://doi.org/10.15587/1729-4061.2021.242304
12. Fortunato, S., Barthélemy, M. (2007). Resolution limit in community detection. Proceedings of the National Academy of Sciences, 104 (1), 36–41. https://doi.org/10.1073/pnas.0605965104
13. Orgnet. Available at: http://www.orgnet.com/
14. Piraveenan, M., Prokopenko, M., Zomaya, A. Y. (2012). On congruity of nodes and assortative information content in complex networks. Networks and Heterogeneous Media, 7 (3), 441–461. https://doi.org/10.3934/nhm.2012.7.441
15. Shergin, V., Udovenko, S., Chala, L. (2020). Assortativity Properties of Barabási-Albert Networks. Data-Centric Business and Applications, 55–66. https://doi.org/10.1007/978-3-030-43070-2_4
16. Shergin, V., Chala, L., Udovenko, S. (2019). Assortativity Properties of Scale-Free Networks. 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), 723–726. https://doi.org/10.1109/picst47496.2019.9061369
17. Shergin, V., Chala, L., Udovenko, S., Pohurska, M. (2018). Assortativity of an elastic network with implicit use of information about nodes degree. CEUR Workshop Proceedings, 131–140. Available at: https://ceur-ws.org/Vol-3018/Paper_12.pdf
18. Noldus, R., Van Mieghem, P. (2015). Assortativity in complex networks. Journal of Complex Networks, 3 (4), 507–542. https://doi.org/10.1093/comnet/cnv005
19. Network data. Available at: https://public.websites.umich.edu/~mejn/netdata/