

OPEN A community partitioning algorithm for cyberspace

Kai Qi, Heng Zhang[✉], Yang Zhou, Yifan Liu & Qingxiang Li

Community partitioning is an effective technique for cyberspace mapping. However, existing community partitioning algorithm only uses the topological structure of the network to divide the community and disregards factors such as real hierarchy, overlap, and directionality of information transmission between communities in cyberspace. Consequently, the traditional community division algorithm is not suitable for dividing cyberspace resources effectively. Based on cyberspace community structure characteristics, this study introduces an algorithm that combines an improved local fitness maximization (LFM) algorithm with the PageRank (PR) algorithm for community partitioning on cyberspace resources, called PR-LFM. First, seed nodes are determined using degree centrality, followed by local community expansion. Nodes belonging to multiple communities undergo further partitioning so that they are retained in the community where they are most important, thus preserving the community's original structure. The experimental data demonstrate good results in the resource division of cyberspace.

As cyberspace continues to develop, the amount of data about various internet resources is also increasing. To better manage, maintain, and optimize cyberspace, it is necessary to conduct an in-depth analysis of these resources and explore their internal connections. If large-scale cyberspace resource data is analyzed using traditional text, numbers, and charts, it is difficult to extract valuable resource information. To better perceive and describe cyberspace resource information, it is divided into communities, and the huge and complex network topology structure is divided into several relatively independent parts. This approach facilitates the management, maintenance, and optimization of the network. This division of resource communities can help us understand the hierarchical structure of the network and determine the important nodes and key paths in the network, allowing us to better optimize network performance, improve network security, and facilitate the location and processing of network faults^{1–4}.

Community partitioning methods may be divided into non-overlapping and overlapping community detection methods, depending on their scope of application. A non-overlapping community only contains nodes belonging exclusively to the community, whereas an overlapping community contains nodes that are simultaneously members of two or more communities⁵. Non-overlapping community detection algorithms mainly include: modularity optimization algorithms^{6–8}, spectral analysis algorithms^{9–11}, information theory algorithms^{12–14}, and label propagation algorithms^{15–17}. When a non-overlapping community detection algorithm assigns a node to a specific community, the node would not be considered for inclusion in other communities. However, real-world cyberspace community structures often have numerous nodes participating in two or more communities. As a result, overlapping community detection algorithms more accurately reflect the community structure of cyberspace^{18–22}. Furthermore, overlapping nodes are often more important for various communities than non-overlapping nodes, which implies that their selection and management can influence the overall quality of the community partition. Overlapping community detection algorithms may be classified as global^{18,23,24} or local information-based community detection algorithms^{25–27}. The global information-based algorithms, while effective, require traversal of the entire network in each iteration. This results in high time complexity and renders them unsuitable for real large-scale networks. To reduce the time complexity of community detection, Lancichinetti et al. proposed the local information-based local fitness maximization (LFM) algorithm²⁶, which introduces a fitness function and a community-size parameter, enabling the detection of overlapping community structures at different levels. Although LFM is very fast and can analyze large-scale networks with millions of nodes, it randomly selects seed nodes for local community expansion during community partitioning. Consequently, it produces a different community partitioning in every run. Due to the differences between network structures and properties, no unique community detection algorithm can be universally adapted to various networks with high accuracy²⁸. Therefore, numerous scholars have carried out a series of studies on different network and purpose needs. Berahmand et al. proposed the Augment Graph Regularization Nonnegative Matrix

Institute of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, Henan, China. [✉]email: 13783651715@163.com

Factorization for Attributed Networks (AGNMF-AN) community partitioning algorithm²⁹. Ma et al. proposed a multi-layer community discovery algorithm based on joint non-negative matrix decomposition³⁰. Zhu et al. proposed a modularity optimization algorithm based on k-plex, which can accurately identify small community structures³¹. Zhe et al. used a greedy maximization modularity algorithm to partition the community of network topology and attribute information³². Bahulkar et al. found the community structure by optimizing the local modularity of the community³³. Zhang et al.'s agglomerative approach used the concept of real connections to discover overlapping communities in the network³⁴.

The common methods used for identifying key nodes in a community include degree centrality³⁵, betweenness centrality³⁶, closeness centrality³⁷, and eigenvector centrality³⁸. However, none of these measures account for the directionality of node-to-node information transfers in the cyberspace community structure. Google introduced the eigenvector centrality-based PageRank (PR) algorithm³⁹, which utilizes the concept of random walks in graph theory¹². PR views the network as a directed graph, with each webpage being a node and each link being a directed edge, and ranks the webpages according to their importance, which is determined from their linkage relations with other webpages.

The traditional community division algorithm only considers the structure of the network and does not consider the attributes of the community structure in the cyberspace. If the traditional community division algorithm is used, it cannot accurately divide the resources in the cyberspace. In a real cyberspace community structure, there are often numerous nodes that exist in two or more communities, and the transmission of information in cyberspace is characterized by direction. Therefore, we propose a community partitioning algorithm of network resource mapping, PR-LFM (PageRank-local Fitness Maximization), which integrates the improved LFM algorithm with PR algorithms. Among them, the LFM algorithm can quickly discover communities based on local information and analyze large-scale networks with up to one million nodes. To better mine the association between nodes and communities, this method simultaneously considers the degree of connection between nodes and the similarity between nodes and communities. The PageRank algorithm not only considers the degree value of nodes, but also the contribution degree of nodes, and it can calculate the ranking of nodes according to the connection relationship of nodes in the network. Furthermore, it exhibits robust performance in the face of large-scale and complex networks. In this study, the PR-LFM algorithm first selects the seed node as the initial community based on the degree central value and then expands the initial community according to the fitness function. Subsequently, by comparing the importance degree of the overlapping community where the overlapping node resides, the overlapping node is retained in the more important community, thereby achieving the same effect as the traditional non-overlapping community division algorithm. Experiments demonstrate that, compared with the traditional non-overlapping community partitioning algorithm, the proposed algorithm takes into account the hierarchy, overlap, and direction of information transmission in the cyberspace and preserves the original community structure in the cyberspace better.

Methods

Cyberspace community structure

Cyberspace is a virtual domain distinct from traditional geographic space, as it is an information network devoid of concepts of distance. Information within it spreads along specific paths and directions, with all kinds of events and processes occurring instantaneously and at zero distance. Hence, cyberspace is virtual, dynamic, directional, and open⁴⁰. The main focus of cyberspace mapping is the community structure of its nodes, which provide insight into the internal workings of cyberspace.

The community structure of cyberspace is characterized by hierarchy, overlaps⁴¹, and directionality. In this context, hierarchy refers to nodes in the network having different levels of organization or structure. For instance, larger communities may contain smaller communities, which could in turn contain even smaller communities⁴². “Overlaps” in the community structure of cyberspace arise from the intersections that exist between different communities in their members, goals, and activities. “Directionality” refers to the fact that information flows from one cyberspace node to another are generally directional. Figure 1 illustrates four overlapping communities. Here,

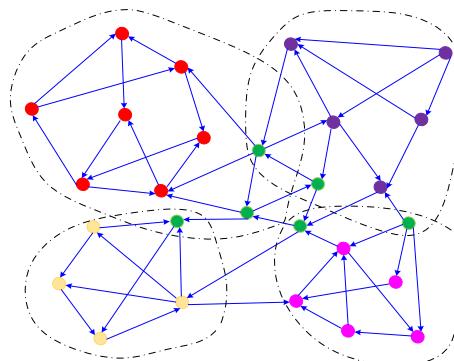


Figure 1. Overlapping of cyberspace communities. Each color represents a different community, and the directed edges represent the direction of information transfer between a pair of nodes. Green nodes are overlapping nodes that are members of two or more communities.

each color represents a different community, and the directed edges represent the direction of information transfer between a pair of nodes. Green nodes are overlapping nodes that are members of two or more communities.

Importance metrics for directed nodes in cyberspace

As information flows in cyberspace are generally directional, we borrowed an idea from Google's PR algorithm, i.e., webpages linked to many high-quality webpages must also be high-quality webpages. In other words, cyberspace nodes shall be evaluated based on the idea wherein nodes linked to important nodes must also be important nodes themselves.

Random walk model for cyberspace

The random walk model is a Markov process with states that randomly move at discrete time intervals. The probability of a state being chosen at each interval is determined by the state transition matrix M . If all nodes connected by a directed edge to a node have equal transition probabilities¹², M is then a matrix of order n :

$$M = [m_{ij}]_{n \times n}. \quad (1)$$

The value of m_{ij} depends on the directed edges between the nodes. If the node j has k outdegrees and the node i is one of its sinks ($i, j = 1, 2, \dots, n$), the value of m_{ij} is $1/k$; otherwise m_{ij} is 0. A random walk model in cyberspace is shown in Fig. 2.

One may thus obtain M , Eq. (2), where each column represents the out node of a node, that is, the links of the current node to other nodes, which sum to 1. Conversely, each row represents the in node of a node, that is, the link from other nodes to the current node.

$$M = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1/3 \\ 1/2 & 0 & 0 & 1/3 \\ 0 & 1 & 0 & 1/3 \end{bmatrix} \quad (2)$$

The probability of the random walk process visiting some node at time t is given by the state distribution of its Markov chain at time t , which can be represented by an n -dimensional column vector R_t . Thus, the probability distribution R_{t+1} for each node being visited at time $t+1$ satisfies

$$R_{t+1} = MR_t. \quad (3)$$

Measurement of node importance in cyberspace using the PR algorithm

The PR algorithm depicts node linkages as a directed graph $A = (V, E)$, where V and E denote the vertices and directed edges of cyberspace, respectively. A random walk model, or a first-order Markov chain, is introduced to represent the process of information transfer between cyberspace nodes. This model can be used to evaluate node importance in cyberspace. The transition matrix of a general random walk model comprises a linear combination of two parts: one is the basic transition matrix M of the directed graph, which assigns equal transfer probabilities from one node to all its sinks, and the other is the random transfer matrix, which has a transition probability of $1/n$ from one node to any other node. This general random walk Markov chain has a stationary distribution, denoted as R . The stationary distribution vector R is defined as the general PR of this directed graph³⁹.

$$R = d.M.R + \frac{1-d}{n}.1 \quad (4)$$

The first term in Eq. (4) represents the probability of each node being visited according to the transfer matrix M , and the second term denotes the completely random probability of each node being visited. d is the linear combination coefficient, and when d is close to 1, the random walk mainly occurs according to the transfer matrix

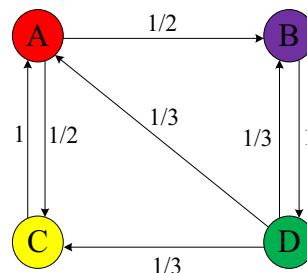


Figure 2. Random walk model for cyberspace. Nodes A, B, C, and D represent different nodes in cyberspace. The directed edges between the nodes represent the direction of linkage in cyberspace, while the weight of each edge represents the probability of random transition between its pair of nodes. For example, if some information is transferred to Node A, its transition probabilities to Nodes B and C are both 1/2. If the information is transferred to Node B, its transition probability to Node D is 1. From Node C, the transition probability to Node A is 1; from Node D, there is a 1/3 probability of transition to one of Nodes A, B, and C.

M. When d is close to 0, the random walk is dominated by random probability. Each component of R is the PR value of each node, which is given by:

$$PR(v_i) = d \left(\sum_{v_j \in M(v_i)} \frac{PR(v_j)}{L(v_j)} \right) + \frac{1-d}{n}, \quad i = 1, 2, \dots, n, \quad (5)$$

where v_i and v_j represent node i and node j , $M(v_i)$ denotes the set of nodes pointing to v_i and $L(v_j)$ denotes the outdegree of node v_j . Table 1 summarizes the commonly used parameters in this paper.

Quality metrics for cyberspace communities

In this study, modularity (Q)⁴³ and normalized mutual information (NMI)⁴⁴ are used as metrics for community structure. Q is a metric that characterizes the degree of connection within a community, i.e., the strength of connectivity among nodes in the community. Based on a large body of experimental evidence, it has been determined that $Q > 0.3$ is indicative of a strong community structure^{45,46}. NMI is a metric used to evaluate the similarity of the calculated clustering solution to the actual community structure, as it measures the clustering similarity of two clustering solutions.

The community partitioning algorithm for cyberspace mapping

In this study, a community partitioning algorithm for cyberspace mapping, known as PR-LFM, was constructed by combining an improved LFM algorithm with PR. This algorithm has three distinct stages: seed node selection, local community expansion, and the partitioning of overlapping nodes. The procedures of the algorithm are as follows:

- Step 1. Seed node selection

In the original LFM algorithm, seed nodes are randomly selected, which results in unstable partitioning results. Given that nodes with high degree centrality are usually key nodes that are important for the dissemination of information, the seed nodes were selected based on the degree centrality. The degree centralities of all nodes in the community are calculated using Eq. (6), and the nodes are then ranked accordingly. The node that has the highest degree centrality is selected as the seed node. This ensures that the same seed nodes are chosen in each run, which assures stable results.

$$DC_i = \frac{k_i}{n-1} \quad (6)$$

- Step 2. Local community expansion

The node with the highest degree centrality from Step (1) is selected as the seed node to start community expansion. The fitness of this community is then calculated using the following equation:

$$f_G = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^\alpha}, \quad (7)$$

where $k_{in}^G(k_{out}^G)$ is the sum of the weights of the edges inside (outside) community G , and α is the resolution parameter that controls the size of the community. The hierarchy of the community structure can be analyzed by selecting different values of α to partition the network (When the α value is low, fewer communities are divided. When the α value is large, more communities are divided).

For any node in the network, a fitness function f_G^i may be defined for the fitness of node i for the community G :

$$f_G^i = f_{G+i} - f_{G-i}, \quad (8)$$

Notation	Explanation
n	Number of network nodes
$M \in \mathbb{R}^{n \times n}$	State transition matrix of network nodes
G	Community
$R^{n \times 1}$	Stationary distribution of the general random walk Markov chain
A	Directed graph
V	Vertices of cyberspace
E	Directed edges of cyberspace
$\mathbf{1}$	Column vector of size n in which all elements are equal to 1

Table 1. Commonly used parameters.

where f_{G-i} and f_{G+i} are the fitnesses of community G before and after node i is added, respectively. The value of f_G^i can be used to determine whether the node joins the community G or not. If $f_G^i > 0$, the addition of node i increases the fitness of community G ; hence, the node should be retained in community G ; otherwise, the node should be removed from community G . The expansion of community G terminates after all nodes have been traversed²⁵. The node with the highest degree of centrality among the unpartitioned nodes is then selected as a seed node for expansion into a new community. This process is repeated until all nodes have been partitioned into at least one community.

- Step 3. Partitioning of overlapping nodes

Based on the communities that were detected in Step (2), if there exists some community _{j} that is entirely a subset of another community _{i} (green and red nodes in Fig. 3, respectively), it is necessary to traverse all communities to eliminate all instances of community _{j} .

Subsequently, the communities that contain overlapping nodes are identified from the remaining communities, and the overlapping nodes are then partitioned into their rightful communities. To this end, the PR value of each overlapping node in each community is calculated, and the node is ultimately retained in the community with the highest PR value. An example is shown in Fig. 4.

Steps (2) and (3) are repeated until all nodes have been partitioned into a community. A flowchart of the algorithm is shown in Fig. 5. In this study, NetworkX⁴⁷, Gephi⁴⁸, and Echarts network analysis packages are used to implement the community detection algorithm.

Experiments and analysis

This study compares the PR-LFM algorithm with the traditional community partitioning algorithm to verify the reliability of the proposed algorithm in small-scale network partitioning. Subsequently, it performs community partitioning based on the real cyberspace router topology data. Experiments demonstrate that the proposed algorithm can retain the real structure of the community better when partitioning the data of large-scale network resources, which can help us better explore the internal connections.

Validation of the PR-LFM algorithm's reliability

To verify the reliability of the PR-LFM algorithm and ascertain its practical value, the Q , NMI, and time of the PR-LFM algorithm are compared with those of other non-overlapping community detection algorithms, such as the GN, Louvain, Infomap, and LPA algorithms. These community detection tests were performed on the classic Karate⁴⁹, Dolphins⁵⁰, Lesmisi⁵¹, Polbooks⁵², and small scale real router networks. The details of these classic networks are presented in Table 2, and the router network information is shown in Table 3.

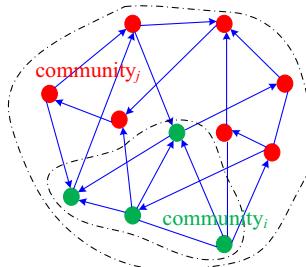


Figure 3. Fully overlapping communities. Nodes in community _{i} exist entirely in community _{j} .

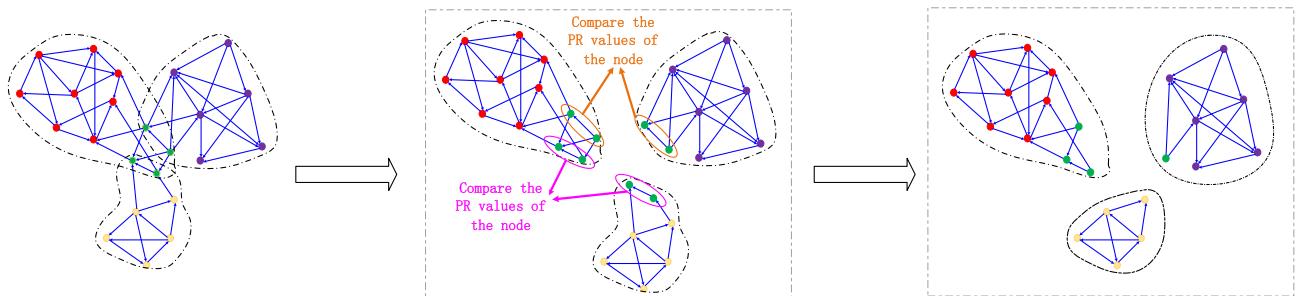


Figure 4. Community detection model for cyberspace mapping. The green overlapping nodes are partitioned into the red, purple, and yellow communities to form a new community structure. Then, the PR value of each overlapping node in each community is calculated, and the node is ultimately retained in the community with the highest PR value.

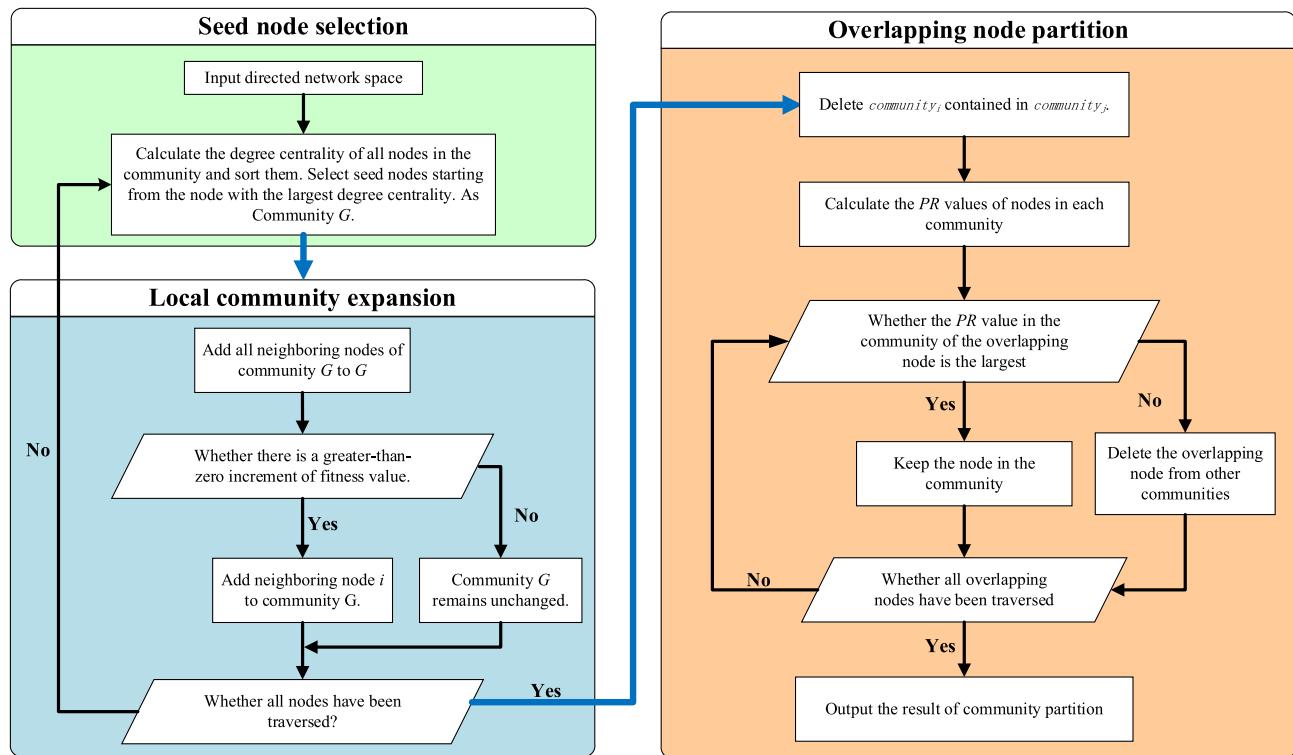


Figure 5. Algorithmic flowchart for the PR-LFM community detection algorithm. The flow chart describes our algorithm in detail, all nodes go through three steps, and the final output node is the partitioned results.

Network name	Nodes	Edges	Community	Average network degree	Map density
Karate	34	78	2	4.588	0.139
Dolphins	62	159	2	3.968	0.084
Lesmis	77	254	–	2.156	0.029
Polbooks	105	441	3	2.752	0.026

Table 2. Classic networks.

	Network I	Network II	Network III	Network IV	Network V	Network VI	Network VII	Network VIII
Nodes	100	143	175	195	200	226	300	350
Edges	288	222	277	285	620	359	906	978

Table 3. Features of router networks.

The Karate network comprises members of a karate club⁴⁹. Our proposed algorithm partitioned this network into four communities and found one overlapping node. The Dolphin network, constructed by Lusseau et al. in New Zealand, models the habits of bottle-nosed dolphins⁵⁰. Our algorithm partitioned this network into seven communities, identifying four overlapping nodes. The Lesmis network represents interactions among characters in Victor Hugo's "Les Misérables."⁵¹ This network was partitioned into four communities, with 12 overlapping nodes. The Polbooks network was constructed by analyzing the political leanings of purchasers of American political books in the Amazon online bookstore⁵². This network was partitioned into three communities, with three overlapping nodes. Finally, the overlapping nodes were retained in the community for which they had the highest PR value, as shown in Fig. 6.

To evaluate the reliability of the algorithm, the Q and NMI of the PR-LFM algorithm on said classic networks were compared to those of other non-overlapping community detection algorithms, as shown in Table 4. It can be observed that the PR-LFM algorithm produces results that are comparable with those of conventional non-overlapping community partitioning algorithms in the aforementioned classic networks. In terms of Q, the PR-LFM algorithm is the best performer on the Karate network, with $Q = 0.4156$. Furthermore, the Q values of the PR-LFM algorithm for the Dolphin network and the Polbooks network are higher than those of the Infomap

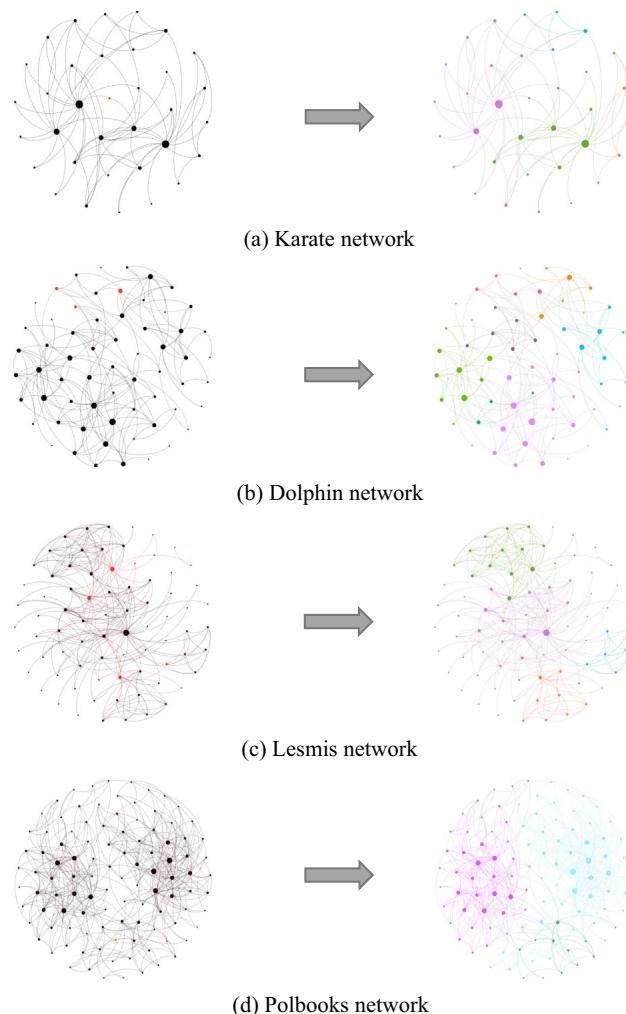


Figure 6. Results of community detection by the PR-LFM algorithm on classic networks. In the figures on the left, the overlapping nodes are colored red. The results of community division are shown on the right, where all nodes in the same community have the same color. The sizes of the nodes are proportional to their degree centrality.

		GN	Louvain	Infomap	LPA	PR-LFM
Karate	Q	0.4013	0.41511	0.35051	0.35470	0.41560
	NMI	0.57983	0.70714	0.56444	0.72096	0.60214
Dolphin	Q	0.51938	0.51958	0.39977	0.45633	0.49830
	NMI	0.55416	0.47425	0.45441	0.46122	0.44658
Lesmis	Q	0.53807	0.55827	0.47092	0.52668	0.51171
	NMI	–	–	–	–	–
Polbooks	Q	0.5168	0.52639	0.39786	0.48114	0.49927
	NMI	0.55845	0.55609	0.46804	0.53410	0.55657

Table 4. NMI and Q values of the PR-LFM algorithm and conventional community detection algorithms on classic networks. Significant values are in [bold].

and LPA algorithms. In terms of NMI, the PR-LFM algorithm is the best performer on the Polbooks network, with $NMI = 0.55657$. Moreover, the NMI of the PR-LFM algorithm on other networks is comparable to those of the conventional algorithms.

Since there is no real community partition result in the router network, the reliability of the proposed algorithm is analyzed in terms of modularity and running time. A comparison of the PR-LFM algorithm proposed

in this study and the traditional non-overlapping community, showing the module degree Q and running time of the algorithm during community partitioning on the router network, is presented in Table 5 and Fig. 7.

As depicted in Fig. 7, when the proposed algorithm divides communities for real router networks, the modularity value consistently surpasses 0.3, indicating that when the proposed algorithm divides communities for router resources in cyberspace, it takes into account the hierarchical and overlapping characteristics of communities in cyberspace and the directivity of transferring resource information, the community structure in cyberspace can still be well preserved. By comparing the modularity with other traditional community partitioning algorithms, we can see that the proposed algorithm can achieve the effect of traditional community partitioning algorithms. Simultaneously, Table 5 reveals that the operation efficiency of the proposed algorithm is significantly higher than that of the GN algorithm. The running time is basically consistent with that of other traditional algorithms, indicating that the algorithm can still be applied to large-scale networks after comprehensive consideration of the characteristics of resource data in the cyberspace.

The experimental results demonstrate that the PR-LFM can achieve the partition effect of the traditional non-overlapping algorithm when classifying resources in the cyberspace. By considering the overlap between communities in real cyberspace, the algorithm proposed in this study can control the number of communities through the parameter α value of community size and consider the direction of resource information transmission in cyberspace through the PageRank algorithm. Compared with the traditional community division algorithm, the algorithm proposed in this study effectively retains the features in the cyberspace. Simultaneously, it achieves a level of efficiency comparable to mainstream algorithm in community division and can be applied to large-scale cyberspace resource data, demonstrating its research value.

Application of the PR-LFM algorithm to cyberspace mapping

Following the effectiveness of the PR-LFM algorithm was validated, a community partitioning test was performed using real router topology data. This dataset contains 4677 router nodes and 6123 edges comprised of child and parent nodes. The experimental procedures are as follows:

- (1) Following multiple trials, it was found that community partitioning can be performed effectively using α values of [0.45–0.54]. Furthermore, the number of detected communities depended on α , which reflects on the hierarchical nature of the community structure. To optimize the algorithm, Q was plotted against α , as shown in the graph of Fig. 8.

Figure 8 illustrates that Q always exceeds 0.3 when $\alpha = [0.45, 0.54]$, indicating that the algorithm produces good results for these values of α . As Q increases significantly when α increases from 0.46 to 0.47, it is implied

	GN	Louvain	Infomap	LPA	PR-LFM
Network I	5.64736	0.01079	0.01248	0.00996	0.01296
Network II	4.55494	0.01496	0.01358	0.01331	0.01392
Network III	15.36345	0.02172	0.01747	0.01894	0.02393
Network IV	12.73726	0.02389	0.02247	0.02046	0.03293
Network V	28.70579	0.02696	0.02745	0.02496	0.03438
Network VI	47.24140	0.02798	0.02897	0.02647	0.03472
Network VII	92.06524	0.04775	0.04893	0.03494	0.05120
Network VIII	128.62424	0.04985	0.05192	0.04742	0.05996

Table 5. Comparison of the running time of router networks on different community partitioning algorithms (seconds).

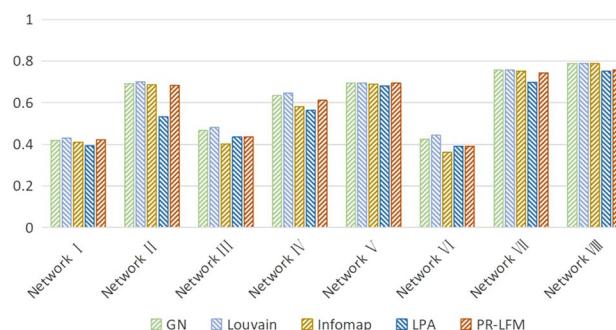


Figure 7. Comparison results of PR-LFM and comparison algorithm Q on the router network.

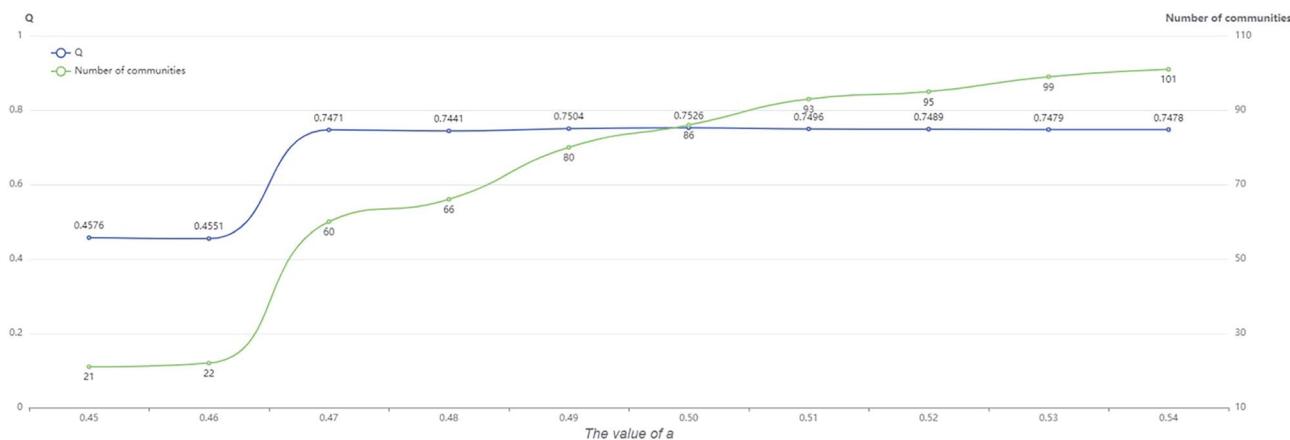


Figure 8. Results of community detection on a real router network using the PR-LFM algorithm with different values of α .

that $\alpha = 0.47$ gives the ideal degree of node connectivity and separation within the detected communities. From $\alpha = 0.47$ to $\alpha = 0.54$, Q only fluctuates slightly around 0.75, which shows that the intra- and inter-community structure was already stable at $\alpha = 0.47$. Q reached its maximum (0.7526) when α was 0.50.

- (2) The degree centralities of all router nodes that correspond to $\alpha = 0.50$ were calculated using Eq. (6), as shown in Fig. 9. Here, ID 1980 has the highest degree centrality, at 0.120615911. Therefore, this node was used as the initial seed node for local community expansion.

The router network was subsequently partitioned based on fitness values, which yielded 86 communities. 80 had overlapping router nodes, and 1359 router nodes, or approximately 29.06% of all router nodes, were present in two or more communities.

- (3) The overlapping router nodes were assigned to their respective communities by comparing their PR values for each community and assigning the node to the community in which their PR was higher. This process was repeated until each node belonged to just one community. The PR-LFM algorithm ultimately partitioned the 4677 router nodes into 86 communities, as shown in Fig. 10.

To test the reliability of PR-LFM on large-scale networks, the PR-LFM algorithm was compared to the aforementioned non-overlapping algorithms in terms of Q (except for the GN algorithm, as it has a time complexity of $O(n^3)$ where n is the number of nodes, making it unsuitable for large-scale networks). Table 6 indicates that the PR-LFM algorithm is second only to the Louvain algorithm in terms of Q . Unlike other algorithms, the PR-LFM algorithm can revisit a node multiple times during community partitioning and directly evaluate each node's contribution to its community, thus enabling the detection of node overlaps. Moreover, the assignment of overlapping nodes to the communities in which they are the most important helps preserve the community structure.

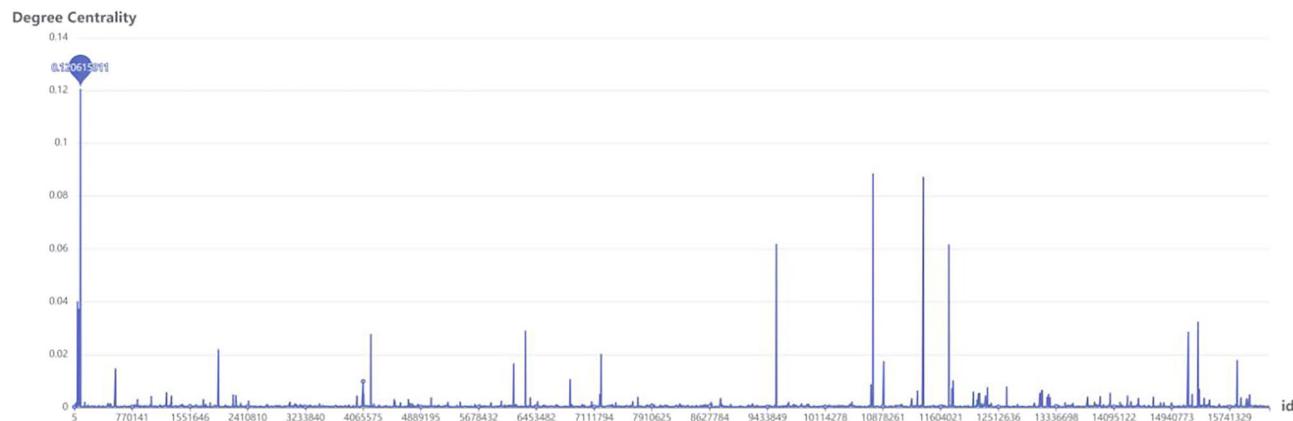


Figure 9. Degree centrality of different router nodes.

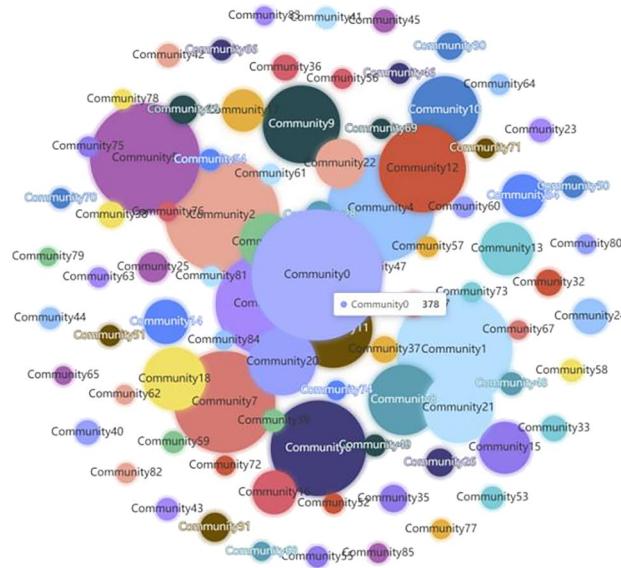


Figure 10. Results of community detection on a real router network using the PR-LFM algorithm. Each circle represents a different community, and the size of the circle represents the number of nodes in the community. For instance, Community₀, which has the most router nodes (378), presents the largest circle.

Algorithm	Q
GN	–
Louvain	0.7758
Infomap	0.6862
LPA	0.6553
PR-LFM	0.7526

Table 6. Q of the PR-LFM algorithm and conventional community detection algorithms on an actual router network. Significant values are in [bold].

Conclusion

To account for the hierarchical and overlapping nature of the community structure of cyberspace and the directionality of its information transfers, we propose the PR-LFM algorithm, a community partitioning method that combines an improved LFM algorithm with the PR algorithm. To account for community overlaps in actual cyberspace, the seed nodes are determined using degree centrality, and the nodes are partitioned into overlapping communities based on a fitness function. To account for the directional nature of information transfer between router nodes, the PR algorithm is used to calculate PR values for overlapping nodes in each of their communities, which are retained in the community where they are the most important (i.e., have the highest PR value). Thus, the overlapping cyberspace communities are partitioned into non-overlapping communities. Experiments were conducted on the classic Karate, Dolphins, Lesmis, and Polbooks networks and real router topology data, which led to the following conclusions:

- (1) Based on comparisons in terms of Q, NMI, and time, the PR-LFM algorithm was found to match the performance of conventional non-overlapping community partitioning algorithms. Furthermore, the PR-LFM algorithm produces stable community partitioning results, as it uses degree centrality to select seed nodes. Therefore, the PR-LFM algorithm is reliable for the partitioning of cyberspace router nodes.
- (2) The PR-LFM algorithm can account for community overlaps during the community partitioning process. When partitioning 4677 router nodes, it detected 1359 overlapping nodes, representing 29.06% of all router nodes. Unlike conventional non-overlapping community detection algorithms, PR-LFM would subsequently partition these overlapping nodes into distinct communities; this helps preserve the community structure.
- (3) The PR-LFM algorithm accounts for the directionality of information transfers between routers in cyberspace by using the PR algorithm to evaluate the importance of each node in the network, which allows the algorithm to accurately characterize router node interactions.
- (4) As the non-overlapping communities derived from this algorithm account for the hierarchy, overlap, and directionality of community structures, it is suitable for mapping large-scale cyber-networks and has wide-

ranging applications in social network analysis, bioinformatics, recommendation systems, and natural language processing.

The goal of this study is to divide the resource node community in cyberspace to manage, maintain, and optimize the mapping data of cyberspace resources. The results demonstrate that the proposed algorithm considers the hierarchy, overlap, and direction of information transmission in the real network structure and can retain the structure in the real cyberspace. Considering that several network resource data have the attributes of geographic space, future research can explore the attributes of network, and geographic space can be combined as the basis for community division. This approach would enable community structure division to consider the characteristics of both network and geographic space, as well as promote the cross-domain application of data and resource information mining. Simultaneously, due to the consideration of community structure characteristics in cyberspace, the proposed algorithm exhibits improved time complexity when compared to the LFM algorithm. Therefore, the following work can focus on optimizing the algorithm and reducing the time complexity to enable efficient handling of large-scale cyberspace resource data.

Data availability

The data used in this study were all derived from citable primary sources, and the actual datasets used in this study may be obtained from the corresponding author.

Received: 20 June 2023; Accepted: 2 November 2023

Published online: 03 November 2023

References

1. Ni, C. C., Lin, Y. Y., Luo, F. & Gao, J. Community detection on networks with Ricci flow. *Sci. Rep.* **9**, 9984 (2019).
2. Kirkley, A. & Newman, M. E. J. Representative community divisions of networks. *Commun. Phys.* **5** (2021).
3. Wang, X. *et al.* Multiscale model construction of elements of point groups in network space by combining Blondel algorithm and the hybrid algorithm of k-core decomposition. *J. Geo-Inf. Sci.* **23**, 2128–2138 (2021).
4. Zhu, Z. *Hierarchical Representation of Network Map* (Wuhan Univ., 2017).
5. Luo, Z. *et al.* New advances in complex network community detection algorithms. *J. Natl. Univ. Def. Technol.* **33**, 47–52 (2011).
6. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7821–7826 (2002).
7. Clauset, & Aaron., Finding local community structure in networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **72**(2), 026132 (2005).
8. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**(23), 8577–8582 (2006).
9. Ng, A., Jordan, M. & Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural J. Inf. Process. Syst.* 849–856 (2002).
10. Donetti, L. & Munoz, M. A. Detecting network communities: A new systematic and efficient algorithm. *J. Stat. Mech. Theory Exp.* **2004**(10), 10012 (2004).
11. Capocci, A., Servedio, V. D. P., Caldarelli, G. & Colaiori, F. Detecting communities in large networks. *Phys. A Stat. Mech. Appl.* (2005).
12. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1118–1123 (2008).
13. Rosvall, M. & Bergstrom, C. T. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci.* **104**(18), 7327–7331 (2007).
14. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**(1), 016110 (2006).
15. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **76**, 036106 (2007).
16. Chen, J. & Saad, Y. Detecting overlapping communities using node-based signal processing. *Phys. Rev. E* **86**(1), 016109 (2012).
17. Li, X., Zhang, W. & Liu, X. Edge based community detection using improved label propagation algorithm. *Phys. A Stat. Mech. Appl.* **442**, 362–374 (2015).
18. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
19. Zhang, S., Wang, R. S. & Zhang, X. S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Phys. A* **374**, 483–490 (2007).
20. Nicosia, V., Mangioni, G., Carchiolo, V. & Malgeri, M. Extending the definition of modularity to directed graphs with overlapping communities. *J. Stat. Mech.* 3166–3168 (2009).
21. Baumes, J., Goldberg, M. & Magdon-Ismail, M. *Efficient Identification of Overlapping Communities* (Springer, 2005).
22. Baumes, J., Goldberg, M. K., Krishnamoorthy, M. S., Magdon-Ismail, M. & Preston, N. Finding communities by clustering a graph into overlapping subgraphs. *Proc. IADIS Int. Conf. Appl. Comput. AC2005* (2005).
23. Shen, H., Cheng, X., Cai, K. & Hu, M. Detect overlapping and hierarchical community structure in networks. *Phys. A* **388**, 1706–1712 (2009).
24. Gregory, S., A fast algorithm to find overlapping communities in networks. *Mach. Learn. Knowl. Discov. Databases*, 408–423 (2008).
25. Schott, Lancichinetti, A., Fortunato, S. & Kertész, J. Detecting the overlapping and hierarchical community structure of complex networks. *New J. Phys.* **11**, 033015 (2009).
26. Lancichinetti, A. & Fortunato, S. Community detection algorithms: A comparative analysis. *Phys. Rev. E* **80**(5), 056117 (2009).
27. Lancichinetti, A., Radicchi, F., Ramasco, J. J. & Fortunato, S. Finding statistically significant communities in networks. *PLoS ONE* **6**(4), e18961 (2008).
28. Rostami, M. *et al.* Community detection algorithms in healthcare applications: A systematic review. *IEEE Access* (2023).
29. Berahmand, K. *et al.* Graph regularized nonnegative matrix factorization for community detection in attributed networks. *IEEE Trans. Netw. Sci. Eng.* **10**(1), 372–385 (2022).
30. Ma, C. *et al.* Identification of multi-layer networks community by fusing nonnegative matrix factorization and topological structural information. *Knowl. Based Syst.* **213**, 106666 (2021).
31. Zhu, J., Chen, B. & Zeng, Y. Community detection based on modularity and k-plexes. *Inf. Sci.* **513**, 127–142 (2020).
32. Zhe, C., Aixin S. & Aokui A. Community detection on large complex attribute network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).
33. Bahulkar, A. *et al.* Community detection with edge augmentation in criminal networks. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. (IEEE, 2018).

34. Zhang, Y. *et al.* True-link clustering through signaling process and subcommunity merge in overlap community detection. *Neural Comput. Appl.* **30**, 3613–3621 (2018).
35. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1**, 215–239 (1978).
36. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40**(1), 35–41 (1977).
37. Sabidussi, G. The centrality of a graph. *Psychometrika* **31**, 581–603 (1966).
38. Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.* **2**, 113–120 (1972).
39. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: Bringing order to the web. Technical report. *Stanford Infolab*, (1999).
40. Ai, T. The development of cartography driven by big data. *Surv. Map Geomat.* **41**, 5–11 (2016).
41. Wang, X. & Liu, Y. A review of community structure algorithms in complex networks. *J. Univ. Electron. Sci. Technol. China* **38**, 537–543 (2009).
42. Blondel, V. D. *et al.* Fast unfolding of communities' hierarchies in large networks. *J. Stat. Mech.* **P10008** (2008).
43. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **69**, 026113 (2004).
44. Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *J. Stat. Mech.* P09008–P09008 (2005).
45. Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **70**, 066111 (2004).
46. Fortunato, S. & Barthélémy, M. Resolution limit in community detection. *Proc. Natl Acad. Sci. U. S. A.* **104**, 36–41 (2007).
47. Hagberg A. A., Schult D. A. & Swart P. J. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, SciPy 200811–16 (2008).
48. Bastian, M., Heymann, S. & Jacomy M. Gephi: An open source software for exploring and manipulating networks[C]. In *Third International AAAI Conference on Weblogs and Social Media* (2009).
49. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
50. Lusseau, D. *et al.* The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **54**, 396–405 (2003).
51. Knuth, D. E. *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, 1993).
52. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, 36–43 (2005).

Acknowledgements

This work is supported by the National Key Research and Development Program of China, No.2016YFB0801301-2, No.20162016YFB0801303.

Author contributions

The principal authors, H.Z. and Y.Z., contributed equally to this work. Y.L. and Q.L. contributed equally in terms of data and theoretical guidance. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023