frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 10 November 2023
DOI 10.3389/fphy.2023.1251319

OPEN ACCESS

EDITED BY
Duxin Chen,
Southeast University, China

REVIEWED BY
Yilun Shang,
Northumbria University, United Kingdom
Salvatore Micciche',
University of Palermo, Italy

*CORRESPONDENCE
Xin-Jian Xu,
xinjxu@shu.edu.cn

RECEIVED 01 July 2023
ACCEPTED 12 October 2023
PUBLISHED 10 November 2023

CITATION
Xu X-J, Chen C and Mendes JFF (2023), Dissimilarity-based hypothesis testing for community detection in heterogeneous networks.
Front. Phys. 11:1251319.
doi: 10.3389/fphy.2023.1251319

Dissimilarity-based hypothesis testing for community detection in heterogeneous networks

Xin-Jian Xu[1]*, Cheng Chen[1] and J. F. F. Mendes[2]

[1]Department of Mathematics, Shanghai University, Shanghai, China, [2]Department of Physics, I3N, University of Aveiro, Aveiro, Portugal

Identifying communities within networks is a crucial and challenging problem with practical implications across various scientific fields. Existing methods often overlook the heterogeneous distribution of nodal degrees or require prior knowledge of the number of communities. To overcome these limitations, we propose an efficient hypothesis test for community detection by quantifying dissimilarities between graphs. Our approach centers around examining the dissimilarity between a given random graph and a null hypothesis which assumes a degree-corrected Erdös–Rényi type. To compare the dissimilarity, we introduce a measure that takes into account the distributions of vertex distances, clustering coefficients, and alpha-centrality. This measure is then utilized in our hypothesis test. To simultaneously uncover the number of communities and their corresponding structures, we develop a two-stage bipartitioning algorithm. This algorithm integrates seamlessly with our hypothesis test and enables the exploration of community organization within the network. Through experiments conducted on both synthetic and real networks, we demonstrate that our method outperforms state-of-the-art approaches in community detection.

KEYWORDS

community detection, stochastic block model, hypothesis test, graph dissimilarity, divergence

1 Introduction

The theory of complex networks has emerged as a powerful tool for studying complex systems. Networks represent interactions between units within a system, with vertices denoting systematic units and edges capturing their interactions [1]. With the increasing availability of real-world data, researchers have been able to conduct studies across various fields. One crucial aspect in these studies is the identification of community structure, where individuals or entities are organized into distinct groups. This task, commonly referred to as community detection [2], shares similarities with graph clustering. Although numerous algorithms have been proposed for community detection including clustering algorithms [3, 4], modularity-based algorithms [5, 6], and dynamic algorithms [7, 8], no single algorithm performs well across all types of networks [9, 10]. Consequently, there is a persistent demand for a general and efficient method for community detection.

From a probabilistic perspective, vertices belonging to the same community are more likely to be connected compared to those in different communities. Therefore, the stochastic block model (SBM) [11] has been widely employed for community detection. The SBM offers a theoretical framework for studying detection thresholds and developing corresponding algorithms. A notable contribution by Decelle et al. [12] introduced the

concept of a phase transition for community detection at the Kesten–Stigum threshold, leading to various investigations into different transition thresholds under varying recovery conditions [13, 14]. Furthermore, numerous algorithms have been proposed for the SBM, often tailored to specific research questions or the characteristics of the system under study. These algorithms encompass spectral methods [15, 16], semi-definite programming methods [17], profile-likelihood maximization [18], and pseudo-likelihood maximization [19]. Of particular interest, Peixoto [20, 21] approached the SBM from a microcanonical perspective, focusing on the number of edges rather than their connection probabilities. This alternative viewpoint offered valuable insights into the SBM.

The standard SBM assumes that vertices within the same community are stochastically equivalent and possess the same expected degree, which does not align with real-world networks where the presence of prominent "hubs" is widespread. To tackle this limitation, Karrer and Newman [22] introduced the degree-corrected SBM (DCSBM) by incorporating vertex-specific "degree parameters" that multiply the edge probability between vertices i and j. Building upon this concept, numerous studies have focused on utilizing the DCSBM for community detection. Zhao et al. [23] established a comprehensive theory for assessing the consistency of community detection in the context of the DCSBM. They also compared various community detection criteria applicable to both the SBM and DCSBM. Chen et al. [24] proposed a method based on convex programming relaxation of modularity maximization and developed a weighted ℓ_1-norm k-medoids algorithm within the DCSBM framework. In contrast, Gao et al. [25] derived the misclassification proportion by evaluating asymptotic minimax risks, which depend on the degree parameter, community size, and connection parameter. It is important to note that all of these algorithms presuppose prior knowledge regarding the number of communities.

In practical scenarios, the only information available to us is the set of vertices and the set of edges, indicating which vertices are connected to each other and which are not. Consequently, determining the appropriate number of communities becomes a challenging task. To the best of our knowledge, existing approaches have primarily focused on the SBM framework. One direction involves initially detecting the optimal community structure for different numbers of communities and then using methods such as minimum description length [26], the Akaike information criterion [27], or the Bayesian information criterion [28] to penalize the model parameters. Another direction involves developing hypothesis tests to determine the number of communities, considering aspects such as asymptotic consistency [29] or the principal eigenvalue of a normalized adjacency matrix [30]. However, both of these approaches suffer from certain limitations. They either require considerable time for large networks or may underestimate or overestimate the number of communities.

The goal of this paper is to simultaneously uncover the number of communities and the corresponding structure in heterogeneous networks in an efficient way. To this end, we propose a novel hypothesis test based on graph dissimilarity, which incorporates three distribution functions of the vertex distance, clustering coefficient, and alpha-centrality. The null hypothesis is assuming that the original network is a one-block DCSBM, i.e., the degree-corrected Erdös–Rényi graph (DCERG), from which one can estimate the connecting parameter and the degree parameter. Then, we compute the dissimilarity between the original network and the posterior DCERG and use the kernel density estimation (KDE) to formulate the dissimilarity distribution among DCERGs generated by the same parameters. If the hypothesis is rejected, we split the network by the bipartitioning algorithm until each subgraph accepts the hypothesis.

2 Hypothesis test

The standard SBM finds its origins in the realms of machine learning and statistics literature. Within theoretical computer science, it is commonly referred to as a planted partition model [31], and in mathematical contexts, it is acknowledged as an inhomogeneous random graph model [32]. This probabilistic generative model for random graphs with community structures seamlessly blends the rigidity of a block model with a stochastic component. It stands as a benchmark in the challenging task of recovering community structures from network data.

To introduce the SBM, we begin with an unweighted and undirected graph denoted as G, consisting of N-labeled vertices that are organized into K blocks. The connections among these K nodes are represented by the adjacency matrix $\boldsymbol{A}$, where $a_{ij} = 1$ if there exists an edge between nodes i and j, and 0 otherwise. It is important to note that self-connections are not allowed, $a_{ii} = 0$. For each node i, we assign a label b_i to represent its membership in a particular community. Consequently, each vertex $i \in [N]$ belongs to a block determined by the prior probability p_j with $j \in [K]$, and these probabilities satisfy the normalization $\sum_{j=1}^{K} p_j = 1$. Additionally, we introduce $\boldsymbol{W}$ as a $K \times K$ matrix, where each element w_{st} represents the probability of connectivity between one vertex in block s and the other in block t. With these definitions, we can now express the conditional expectation of the adjacency matrix $\boldsymbol{A}$ given the block assignments $\boldsymbol{b}$ as follows:

$$\mathrm{E}\left(a_{ij}|\boldsymbol{b}\right) = w_{b_i,b_j}. \tag{1}$$

When all labels are identical, the model simplifies to the classic Erdös–Rényi graph (ERG) [33], where meaningful reconstruction of communities becomes unfeasible. In the context of real-world networks, the model can be adjusted by maximizing this expectation concerning vertex labels $\boldsymbol{b}$. The primary objective of the community detection problem is to accurately reconstruct these labels.

In the standard SBM, the connecting probabilities between any two vertices within the same block are uniform. In such a configuration, the emergence of "hubs" becomes unlikely, and maximizing the log-likelihood function based on it tends to partition the graph into two groups: one consisting of high-degree vertices and the other composed of low-degree vertices. To overcome this limitation, Karrer and Newman [22] introduced the DCSBM, which replaces Eq. 1 with the following equation:

$$\mathrm{E}\left(a_{ij}|\theta,\boldsymbol{b}\right) = \theta_i\theta_j w_{b_i,b_j}, \tag{2}$$

where θ_i is a degree parameter. In contrast to the SBM, the DCSBM modifies the edge probability between vertices i and j by multiplying it with the product of $\theta_i\theta_j$. Notably, the DCSBM simplifies to the standard SBM when $\theta_i = 1$ is the same for every vertex $i \in [N]$. The value of θ_i plays a crucial role in determining the degree of vertex i, enabling flexibility in accommodating arbitrary degree variations within blocks. However, it is essential to ensure that θ_i satisfies specific constraints. In this paper, we impose the constraint that $\sum \theta_i \delta_{b_i,s} = 1$ holds true for all blocks.

A challenge encountered in both the SBM and DCSBM is the necessity of prior knowledge regarding the precise number of blocks in the network. However, the use of hypothesis testing offers a potential solution to mitigate this requirement. Essentially, the task of determining whether a DCSBM consists of either K or $K + 1$ blocks can be viewed as an inductive decision between one block or two. This line of thinking leads us to the formulation of a null hypothesis: the network follows a one-block DCSBM, i.e., the DCERG. The expected adjacency matrix for the DCERG is expressed as follows:

$$\mathrm{E}(\boldsymbol{A}) = \boldsymbol{DZD}, \tag{3}$$

with $\boldsymbol{D} = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_N)$ and $\boldsymbol{Z} = Nw\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} - w\boldsymbol{I}$, where $\boldsymbol{e}$ is a vector with $e_i = 1/\sqrt{N}$ for $i \in [N]$ and $\boldsymbol{I}$ is the identity matrix. Assuming that the graph is generated by the DCERG, we need to estimate θ and w. The former is given by the following equation:

$$\hat{\theta}_i = \frac{k_i}{\sum_{i=1}^{N} k_i}, \tag{4}$$

with $k_i = \sum_{j=1}^{N} a_{ij}$ being the degree of vertex i, while the later can be written as follows

$$\hat{w} = \frac{\sum_{i=1,j=1}^{N} e_{ij}}{N(N-1)} \tag{5}$$

with $e_{ij} = \theta_i^{-1} a_{ij} \theta_j^{-1}$. Now, the problem becomes to distinguish the DCSBM(N, p, W, θ) and DCERG$(N, \hat{w}, \hat{\theta})$. If they demonstrate a significant dissimilarity, we reject the null hypothesis and partition the community. This process continues until each subgraph conforms to a DCERG, allowing us to determine the number of communities in the network simultaneously.

In the realm of graph analysis, gauging the structural dissimilarity of large graphs presents a formidable challenge due to the frequently unwieldy computational complexity associated with analysis techniques [34, 35]. Despite the abundance of literature on this subject, the majority of studies have traditionally focused on examining simple graphs, often overlooking factors such as degree heterogeneity and community structure. To surmount this limitation, Xu et al. [36] introduced a precise and efficient method for quantifying dissimilarities between graphs, denoted as G and G'. Their approach adopts a perspective rooted in probability distribution functions:

$$\begin{aligned} D(G, G') = {} & \gamma_1 \sqrt{\frac{\mathcal{J}(Q_l(G), Q_l(G'))}{\log 2}} + \gamma_2 \sqrt{\frac{\mathcal{J}(Q_c(G), Q_c(G'))}{\log 2}} \\ & + \gamma_3 \sqrt{\frac{\mathcal{J}(Q_\alpha(G), Q_\alpha(G'))}{\log 2}}, \end{aligned} \tag{6}$$

where γ_1, γ_2, and γ_3 are positive constants satisfying $\gamma_1 + \gamma_2 + \gamma_3 = 1$. The values of these three parameters reflect the influence of global (first term), local (second term) features, and heterogeneity (third term) on the dissimilarity measure. $Q_l(G) = \{q_l(i)\} = \{\sum_{i=1}^{N} n_{ik}/N(N-1)\}$ denotes the average distance distribution, and n_{ik} is the number of vertices at distance k from vertex i. $Q_c(G) = \{q_c(i)\} = \{[\pi_c; N - \sum_{i=1}^{N} \pi_c(i)]/N\}$ represents the average clustering coefficient distribution, and π_c is the clustering coefficient of vertex i in an increasing order. $Q_\alpha(G) = \{q_\alpha(i)\} = \{[\pi_\alpha; N - \sum_{i=1}^{N} \pi_\alpha(i)]/N\}$ corresponds to the average centrality distribution, and π_α is the α-centrality of vertex i in an increasing order. $\mathcal{J}(\boldsymbol{q}_1, \boldsymbol{q}_2) = \frac{1}{2}\sum_i q_1(i) \ln[2q_1(i)/(q_1(i) + q_2(i))] + \frac{1}{2}\sum_i q_2(i) \ln[2q_2(i)/\sum_i (q_1(i) + q_2(i))]$ is the Jensen–Shannon divergence. Defined in this way, D captures both global and local dissimilarities of the two graphs. Moreover, it is easy to confirm that $D \in [0, 1)$. Looking at Eq. 6, it becomes evident that D is a random variable. To derive the probability distribution of D, we employ the KDE technique [37]. Given a collection of samples D_1, D_2, ..., D_n, the KDE offers a means to estimate the distribution as follows:

$$P(D) = \frac{1}{n} \sum_{i=1}^{n} \kappa(D - D_i), \tag{7}$$

where $\kappa(D - D_i) = e^{-\frac{\|D - D_i\|^2}{2\sigma^2}}$ and σ is the bandwidth parameter to control the smoothness of the estimate. In the present work, we set $\sigma = 0.34$. Finally, we can calculate the p-value to accept or reject the null hypothesis.

Building upon the aforementioned rationale, we propose a two-stage hypothesis testing algorithm (see Algorithm 1). In the first stage, we employ hypothesis testing to ascertain if the network is a single-community network, specifically a DCERG. The detailed procedure is outlined as follows: i) We begin by assuming that the target network G adheres to the DCERG and proceed to estimate its degree parameter θ and edge parameter w. ii) Utilizing the estimated $\hat{\theta}$ and $\hat{w}$, we generate n DCERGs denoted as G_1, G_2, ..., G_n. Subsequently, we compute the dissimilarities $D(G, G_i)$ and $D(G_i, G_j)$, where i and j are distinct and range from 1 to n. iii) Employing the KDE, we estimate the dissimilarity distribution P between isomorphic single-community networks. iv) We employ the disparity in average dissimilarity between the target network G and all generated DCERGs, denoted as $\bar{D}$, as the test statistic. Subsequently, we utilize the dissimilarity distribution P of isomorphic single-community networks as the test distribution in the application of hypothesis testing to ascertain whether G constitutes a single-community network.

```
A ← adjacency matrix of G
θ̂_i ← k_i / Σ_{i=1}^{N} k_i,  ŵ ← Σ_{i=1,j=1}^{N} E_ij / N(N−1)
For i = 1, 2, ..., 50
    G_i ← DCERG(N, ŵ, θ̂)
    D̄ = Σ D(G, G_i)/50
For all i ≠ j
    D_ij ← D(G_i, G_j)
P̂(D(G, DCERG(N, ŵ, θ̂))) ← KDE(D_ij)
pval ← P̂(D̄ > D)
If pval < significant level α
    i) For each edge e_ij
        compute the edge betweenness B_ij
        compute the edge clustering coefficient C_ij
        L_ij ← β1 B_ij − β2 C_ij
        remove edge e_ij with L = max(L_ij)
    ii) If the graph is connected
        go back to i)
       Else
        Output G′, G″
       End if
Else
    Output G
End if
```

Algorithm 1. Hypothesis test algorithm.

If the null hypothesis is rejected, the algorithm progresses to the second stage, where the original target network undergoes division into two distinct networks. In this phase, we enhance the Newman–Girvan algorithm [18] by taking into consideration the significance of edges with regards to network connectivity and the local clustering of edges within the network. This approach simultaneously incorporates both global and local information, thereby augmenting the precision of community detection. The specific procedure is delineated as follows: i) computing the edge betweenness B_{ij} and the edge clustering coefficient C_{ij} for the target network G; ii) defining $L_{ij} = \beta_1 B_{ij} - \beta_2 C_{ij}$, where the values of β_1 and β_2 represent the balance between connectivity and clustering, and eliminating edge E_{ij} associated with the maximum L_{ij}; and iii) cycling back to step i) and iterating until the network is no longer connected. Consequently, the binary partitioning algorithm yields the original G as two separate networks, denoted as G' and G''. However, following the split, these two networks may not necessarily conform to the single-community assumption. Consequently, the testing algorithm (Stage I) and partitioning algorithm (Stage II) are iteratively applied until each network ultimately embodies a single-community network.

The computational efficiency of our algorithm is characterized by its polynomial time. Initially, the generation of an N-node DCERG incurs a cost of $O(N)$. The computation of dissimilarity relies on the shortest path, a process that can be efficiently implemented in $O(M + NlogN)$ through the use of Fibonacci heaps. In the present work, we generate $C_{50}^2 = 1225$ dissimilarities. Generalizing to any value m, the generation of C_m^2 dissimilarities results in a time complexity of $O(m^2)$, influencing the overall cost of the KDE. During the edge removal of partitioning, the time complexity associated with edge betweenness is $O(NM)$, while the time complexity of edge clustering is $O(N + M)$. Finally, for a K-communities network, the algorithm iterates $k - 1$ times, contributing to the overall efficiency of the approach.

3 Application to block models

To assess the effectiveness of our algorithm, we initiate testing on the balanced DCSBM, where each block is of identical size. In particular, we fixed the parameters at $N = 1000$, $K = 2$, and $w_{11} = w_{22} = 0.2$. The degree parameters, denoted as θ_i, are drawn from an adjusted normal distribution characterized by $\theta \sim (|\text{Normal}(0, 0.25)| + 1 - \frac{1}{\sqrt{2\pi}})$, which exhibits a right-skewed profile. It should be noted that we also explore alternative distributions, although those results are not presented here. To maintain generality, we set the mean of this distribution to $E(\theta) = 1$. The process of generating the graph aligns with a straightforward implementation of the block model. It involves (i) drawing a Poisson-distributed number of edges between each pair of blocks 1 and 2 with $w_{12} = w_{21}$ (or $w_{11}/2 = w_{22}/2$ for intra-block connections and (ii) probabilistically assigning each end of an edge to a vertex within the respective block, guided by the parameter θ_i.

To explore different levels of community structure within the generated networks, we systematically increased the value of $w_{12}(=w_{21})$ from 0.02 to 0.2 in increments of 0.02. We calculate the error bars on p-values based on the outcomes of 100 random runs. Essentially, a larger p-value suggests that the hypothesis test perceives the graph as being closer to an ERG. As illustrated in Figure 1A, we observed an increasing trend in the p-value as w_{12} increases, indicating a diminishing block structure in the network. Figure 1B provides a visual representation of the adjacency matrix for the case of $w_{12} = 0.02$. In this representation, rows and columns are ordered based on the underlying community structure. Importantly, the block structure detected by our algorithm closely aligns with the intended model settings.

We proceed by applying our algorithm to the DCSBM with unbalanced blocks. Specifically, we examine the scenario where the two blocks have different sizes, denoted as n_1 and n_2, respectively. To investigate the impact of community size, we set $w_{12} = w_{21} = 0.02$ and $w_{11} = w_{22} = 0.2$.

Figure 2A illustrates the behavior of the p-value as n_1 increases from 50 to 100. Notably, the p-value consistently decreases with the growth of n_1. This trend is straightforward to comprehend as the detection of the planted block becomes increasingly easier with a larger n_1. In fact, the DCSBM demonstrates a clear block structure when $n_1 \geq 77$. In contrast, in Figure 2B, we set $n_1 = 100$ and plot the p-value against the varying values of w_{12}. Here, an interesting observation emerges: the p-value displays a consistent rise with an increase in w_{12}. This outcome aligns with expectations since the graph gradually loses its block structure, particularly noticeable when $w_{12} \geq 0.068$.

4 Application to empirical networks

We now apply our algorithm to real-world networks. The first example we consider is the network of a karate club at an American university. This network consists of 34 nodes, and the relationships between these nodes were recorded by Zachary [38] over a span of 2 years. Due to a disagreement between an instructor (node 0) and an administrator (node 33) regarding class fees, the club ultimately split into two distinct groups. The knowledge of the members within each group makes the karate club network an ideal benchmark for studying community detection.

Upon applying our algorithm to this network, the results obtained are shown in Figure 3A. In the figure, solid circles and squares represent clusters corresponding to instructors and administrators, respectively. Overall, our algorithm successfully splits the vertices in accordance with the known communities, aside from a misclassification of two vertices (nodes 8 and 9) located on the boundary between the two groups. Furthermore, Figure 3B presents a density image of the adjacency matrix, serving as additional confirmation of the block structure within the network.

As a second real-world example, we turn our attention to the American college football network [39]. This network is comprised of teams within a league, with each node representing an individual team. Nodes are connected if the corresponding teams played against each other during a specific season. Specifically, our dataset focuses on the 2000 season of the American College Football Division 1-A and includes a total of 115 teams. These teams are organized into 12 conferences, and it is worth noting that games are more commonly played between members of the same conference rather than between teams from different conferences, resulting in a recognizable community structure.

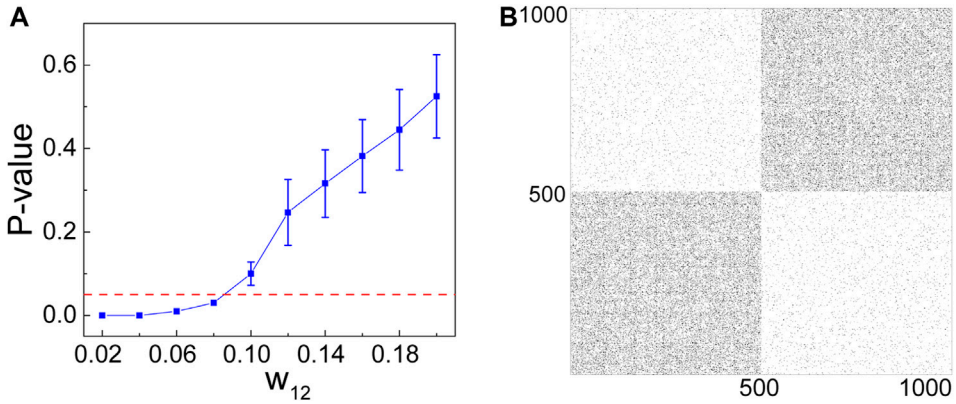

FIGURE 1
Simulation results of the hypothesis test algorithm for the balanced two-block DCSBM: p-value as a function of the connecting parameter w_{12} **(A)** and the illustration of the adjacency matrix for w_{12} = 0.02 **(B)**. Dashed line corresponds to the significant level α = 0.05.

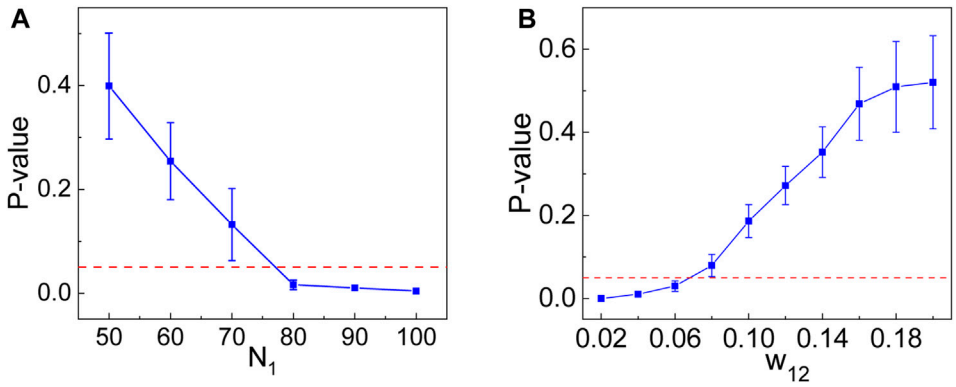

FIGURE 2
p-value as a function of n_1 **(A)** and w_{12} **(B)** for the unbalanced two-block DCSBM. The dashed lines correspond to the significant level α = 0.05.

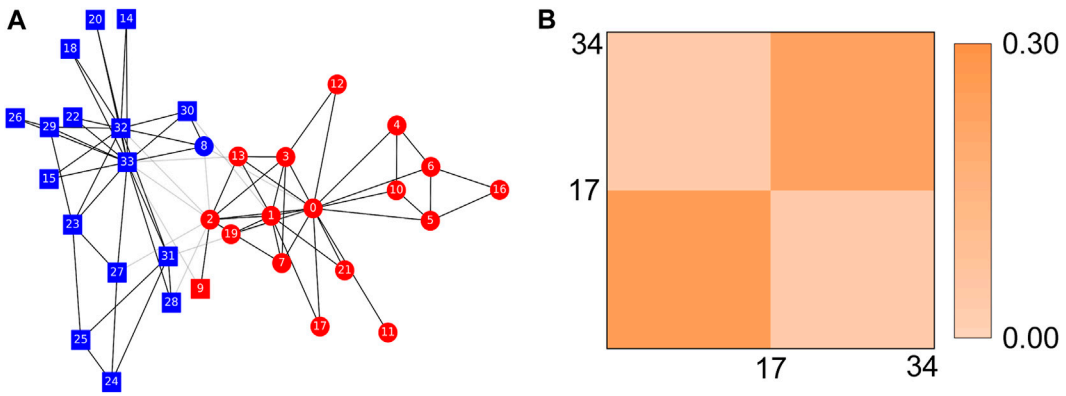

FIGURE 3
Performance of the hypothesis test algorithm for the karate club: the illustration of the community division **(A)** and the density plot for the network **(B)**.

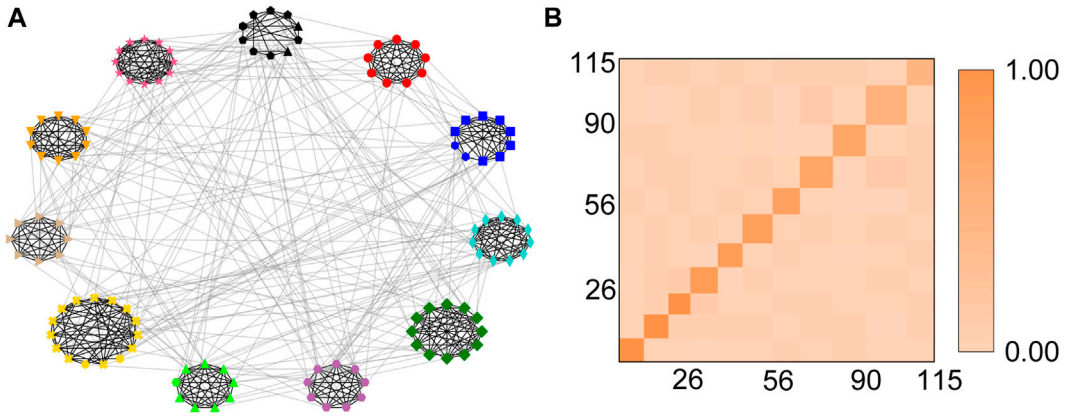

FIGURE 4
Community division **(A)** and density matrix **(B)** for the American college football network.

TABLE 1 Comparison of the results of the hypothesis test algorithm to the ground truth and those of the state-of-the-art algorithms.

	Karate club			College football		
	Communities	S_{AR}	F_1	Communities	S_{AR}	F_1
Hypothesis test	2	0.7717	0.9410	11	0.8927	0.8697
Motif-based k-means	2	0.6682	0.9117	10	0.7939	0.8120
Modularity-based	3	0.5684	0.5189	6	0.4741	0.3711
Louvain	4	0.4646	0.3033	10	0.8035	0.6961
Infomap	3	0.5906	0.5666	10	0.8165	0.6940

In Figure 4A, we present the community structure obtained through the application of our algorithm to this network. This analysis reveals that the majority of teams have been accurately grouped with other teams from their respective conferences. However, there are a few independent teams that have been assigned to conferences with which they share the closest associations, demonstrating a high level of agreement between the algorithm's results and the ground truth community structure. Furthermore, Figure 4B displays a density plot of the adjacency matrix, providing further clarity on this phenomenon.

To quantitatively compare the results of our algorithm to the ground truth and those of the state-of-the-art methods, we introduce the following two measures: the adjusted Rand index S_{AR} and F_1 score. Given two kinds of classifications P_a and P_b, we denote the count of node pairs that classified together in both partitions by q_{11}, classified together in P_a but different in P_b by q_{10}, different in P_a but classified together in P_b by q_{01}, and different in both by q_{00}. It is worth noting that $w_{11} + w_{10} + w_{01} + w_{00} = C_n^2 = M$, and the adjusted Rand index is defined as follows [40]:

$$S_{AR} = \frac{w_{11} - \frac{1}{M}(w_{11} + w_{10})(w_{11} + w_{01})}{\frac{1}{2}[(w_{11} + w_{10}) + (w_{11} + w_{01})] - \frac{1}{M}(w_{11} + w_{10})(w_{11} + w_{01})}. \tag{8}$$

Another measure comparing P_a and P_b is F_1 score, defined as follows [41]:

$$F_1 = \frac{2\text{precision}(P_a, P_b)\text{recall}(P_a, P_b)}{\text{precision}(P_a, P_b) + \text{recall}(P_a, P_b)}, \tag{9}$$

with $\text{precision}(P_a, P_b) = |P_a \cap P_b|/|P_b|$ and $\text{recall}(P_a, P_b) = |P_a \cap P_b|/|P_a|$.

As depicted in Table 1, our method outperforms state-of-the-art approaches in identifying communities for both real networks. Specifically, we successfully identified two communities in the karate club network and 11 communities in the football network, surpassing the results obtained by other methods. Notably, our approach yielded the highest values for S_{AR} and F_1, indicating a superior alignment with the ground truth communities.

5 Conclusion

As a prominent model for the analysis of structural data, the SBM and its variants, particularly the DCSBM, have garnered significant attention in the realm of community detection within networks [42]. The DCSBM, in particular, stands out for its efficacy in handling networks characterized by a highly skewed degree distribution. In this paper, we introduced a novel hypothesis test designed for community detection in complex networks, making dual contributions in terms of both model and algorithm. On the modeling front, we introduced a graph dissimilarity measure that incorporates the vertex distance distribution, clustering coefficient distribution, and alpha-centrality distribution. Utilizing this dissimilarity measure between the DCSBM and DCERG,

we proposed a hypothesis testing statistic. In the algorithmic domain, we devised a two-stage algorithm. Initially, we determined whether the original network adhered to the DCERG. If not, we iteratively bipartitioned it until each subgraph conformed to the DCERG. A new criterion for bipartitioning was introduced, integrating edge betweenness and edge clustering coefficient. We applied the algorithm to both synthetic and real networks. Overall, the proposed method marks a significant advancement over existing state-of-the-art approaches. It demonstrates feasibility in detecting communities within networks characterized by broad degree distributions, even when the actual number of communities is unknown.

There are several promising directions for future research in this field. One key area involves exploring alternative approaches to measuring graph dissimilarity, as it remains an open problem. Particularly, for networks with higher-order architecture, it would be beneficial to consider measures that go beyond pairwise interactions to enhance the model's capacity [43]. Additionally, while the Gaussian distribution is commonly chosen for the kernel density distribution, it may be valuable to explore other distributions, like the widely used Epanechnikov distribution in financial data analysis, to cater to specific interests. In terms of computational complexity, a crucial avenue would involve determining the theoretical distribution for dissimilarity, ultimately contributing to significant reductions in computational overhead. Furthermore, the proposed framework can be extended to incorporate more sophisticated block models, such as exponential [44], multilevel [45], and dynamic [46] models, offering additional benefits and insights.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

X-JX and JM conceived and designed the study. CC developed the code and performed the simulations. X-JX and JM interpreted the results and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

X-JX and CC acknowledge financial support from NSFC 12071281 and STCSM 22JC1401401. JM acknowledges financial support from the project I3N and FCT/MEC UIDB/50025/2020 and UIDP/50025/2020.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Barabási A-L. *Network science*. Cambridge: Cambridge University Press (2015).

2. Fortunato S. Community detection in graphs. *Phys Rep* (2010) 486:75–174. doi:10.1016/j.physrep.2009.11.002

3. Maqbool O, Babri HA. The weighted combined algorithm: a linkage algorithm for software clustering. In: *Proceedings of the 8th European conference on software maintenance and reengineering* (2004). p. 15–24.

4. Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci* (2006) 103(23):8577–82. doi:10.1073/pnas.0601602103

5. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Phys Rev E* (2004) 70:066111. doi:10.1103/physreve.70.066111

6. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech* (2008) 2008:P10008. doi:10.1088/1742-5468/2008/10/p10008

7. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* (2005) 435:814–8. doi:10.1038/nature03607

8. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci* (2008) 105(4):1118–23. doi:10.1073/pnas.0706851105

9. Hric D, Darst RK, Fortunato S. Community detection in networks: structural communities versus ground truth. *Phys Rev E* (2014) 90:062805. doi:10.1103/physreve.90.062805

10. Yang Z, Algesheimer R, Tessone CJ. A comparative analysis of community detection algorithms on artificial networks. *Sci Rep* (2016) 6:30750. doi:10.1038/srep30750

11. Holland PW, Laskey KB, Leinhardt S. Stochastic blockmodels: first steps. *Soc Netw* (1983) 5:109–37. doi:10.1016/0378-8733(83)90021-7

12. Decelle A, Krzakala F, Moore C, Zdeborová L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys Rev E* (2011) 84:066106. doi:10.1103/physreve.84.066106

13. Abbe E, Sandon C. Community detection in general stochastic block models: fundamental limits and efficient algorithms for recovery. In: *Proceedings of the 56th annual symposium on foundations of computer science* (2015). p. 670–88.

14. Abbe E, Bandeira A, Hall G. Exact recovery in the stochastic block model. *IEEE Trans Inform Theor* (2016) 62(1):471–87. doi:10.1109/tit.2015.2490670

15. Rohe K, Chatterjee S, Yu B. Spectral clustering and the high-dimensional stochastic block model. *Ann Statist* (2011) 39(4):1878–915. doi:10.1214/11-aos887

16. Sarkar P, Bickel P. Role of normalization in spectral clustering for stochastic blockmodels. *Ann Statist* (2015) 43(3):962–90. doi:10.1214/14-aos1285

17. Guédon O, Vershynin R. Community detection in sparse networks via grothendieck's inequality. *Probab Theor Relat Fields* (2016) 165:1025–49. doi:10.1007/s00440-015-0659-z

18. Bickel PJ, Chen A. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc Natl Acad Sci* (2009) 106(50):21068–73. doi:10.1073/pnas.0907096106

19. Amini AA, Chen A, Bickel PJ, Levina E. Pseudo likelihood methods for community detection in large sparse networks. *Ann Statist* (2013) 41(4):2097–122. doi:10.1214/13-aos1138

20. Peixoto TP. Parsimonious module inference in large networks. *Phys Rev Lett* (2012) 110:148701. doi:10.1103/physrevlett.110.148701

21. Peixoto TP. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Phys Rev X* (2015) 5:011033. doi:10.1103/physrevx.5.011033

22. Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. *Phys Rev E* (2011) 83:016107. doi:10.1103/physreve.83.016107

23. Zhao Y, Levina E, Zhu J. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann Statist* (2012) 40(4):2266–92. doi:10.1214/12-aos1036

24. Chen Y, Li X, Xu J. Convexified modularity maximization for degree-corrected stochastic block models. *Ann Statist* (2018) 46(4):1573–602. doi:10.1214/17-aos1595

25. Gao C, Ma Z, Zhang AY, Zhou HH. Community detection in degree-corrected block models. *Ann Statist* (2018) 46(5):2153–85. doi:10.1214/17-aos1615

26. Rosvall M, Bergstrom CT. An information-theoretic framework for resolving community structure in complex networks. *Proc Natl Acad Sci* (2007) 104(18):7327–31. doi:10.1073/pnas.0611034104

27. Burnham KP, Anderson DR. *Model selection and multi-model inference: a practical information-theoric approach*. Colorado: Springer-Verlag (2004).

28. Handcock MS, Raftery AE, Tantrum JM. Model-based clustering for social networks. *J Roy Stat Soc A* (2007) 170:301–54. doi:10.1111/j.1467-985x.2007.00471.x

29. Zhao Y, Levina E, Zhu J. Community extraction for social networks. *Proc Natl Acad Sci* (2011) 108(18):7321–6. doi:10.1073/pnas.1006642108

30. Bickel PJ, Sarkar P. Hypothesis testing for automated community detection in networks. *J Roy Stat Soc B* (2016) 78:253–73. doi:10.1111/rssb.12117

31. Bui T, Chaudhuri S, Leighton F, Sipser M. Graph bisection algorithms with good average case behavior. *Combinatorica* (1987) 7:171–91. doi:10.1007/bf02579448

32. Shang Y. Characterization of expansion-related properties of modular graphs. *Disc Appl Math* (2023) 338:135–44. doi:10.1016/j.dam.2023.06.002

33. Erdös P, Rényi A. On the evolution of random graphs. *Publ Math Inst Hung Acad* (1960) 5:17–61.

34. Emmert-Streib F, Dehmer M, Shi Y. Fifty years of graph matching, network alignment and network comparison. *Inform Sci* (2016) 346–347:180–97. doi:10.1016/j.ins.2016.01.074

35. Schieber TA, Carpi L, Díaz-Guilera A, Pardalos PM, Masoller C, Ravetti MG. Quantification of network structural dissimilarities. *Nat Commun* (2017) 8:13928. doi:10.1038/ncomms13928

36. Xu X-J, Chen C, Mendes JFF. Quantifying dissimilarities between heterogeneous networks with community structure. *Physica A* (2022) 588:126574. doi:10.1016/j.physa.2021.126574

37. Parzen E. On estimation of a probability density function and mode. *Ann Math Stat* (1962) 33:1065–76. doi:10.1214/aoms/1177704472

38. Zachary W. An information flow model for conflict and fission in small groups. *J Anthropol Res* (1977) 33:452–73. doi:10.1086/jar.33.4.3629752

39. Newman MEJ. Communities, modules and large-scale structure in networks. *Nat Phys* (2012) 8:25–31. doi:10.1038/nphys2162

40. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: *Proceedings of the 26th annual international conference on machine learning* (2010). p. 1073–80.

41. Larsen B, Aone C. Fast and effective text mining using linear-time document clustering. In: *Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining* (1999). p. 16–22.

42. Snijders TAB, Nowicki K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J Classification* (1997) 14(1):75–100. doi:10.1007/s003579900004

43. Lacasa L, Stramaglia S, Marinazzo D. Beyond pairwise network similarity: exploring mediation and suppression between networks. *Commun Phys* (2021) 4:136. doi:10.1038/s42005-021-00638-9

44. Schweinberger M. Consistent structure estimation of exponential-family random graph models with block structure. *Bernoulli* (2020) 26(2):1205–33. doi:10.3150/19-bej1153

45. Chabert-Liddell S-C, Barbillon P, Donnet S, Lazega E. A stochastic block model approach for the analysis of multilevel networks: an application to the sociology of organizations. *Comput Statist Data Anal* (2021) 158:107179. doi:10.1016/j.csda.2021.107179

46. Bartolucci F, Pandolfi S. An exact algorithm for time-dependent variational inference for the dynamic stochastic block model. *Pattern Recognit Lett* (2020) 138:362–9. doi:10.1016/j.patrec.2020.07.014