Hybrid-order Stochastic Block Model

Xunxun Wu,[1] **Chang-Dong Wang,** [1,2,3*] **Pengfei Jiao** [4]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] Guangdong Province Key Laboratory of Computational Science, Guangzhou, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4] Center of Biosafety Research and Strategy, Law school of Tianjin University, Tianjin, China
xunxunwu99@outlook.com, changdongwang@hotmail.com, pjiao@tju.edu.cn

Abstract

Community detection is a research hotspot in machine learning and data mining. However, most of the existing community detection methods only rely on the lower-order connectivity patterns, while ignoring the higher-order connectivity patterns, and unable to capture the building blocks of the complex network. In recent years, some community detection methods based on higher-order structures have been developed, but they mainly focus on the motif network composed of higher-order structures, which violate the original lower-order topological structure and are affected by the fragmentation issue, resulting in the deviation of community detection results. Therefore, there is still a lack of community detection methods that can effectively utilize higher-order connectivity patterns and lower-order connectivity patterns. To overcome the above limitations, this paper proposes the Hybrid-order Stochastic Block Model (HSBM) from the perspective of the generative model. Based on the classical stochastic block model, the generation of lower-order structure and higher-order structure of the network is modeled uniformly, and the original topological properties of the network are maintained while using higher-order connectivity patterns. At the same time, a heuristic algorithm for community detection is proposed to optimize the objective function. Extensive experiments on six real-world datasets show that the proposed method outperforms the existing approaches.

Introduction

Community detection is a research hotspot in network analysis that aims to divide the network into substructures with tight internal connections and sparse external connections, and plays an important role in various fields such as human social interaction, economy and trade, biological information, transportation and electricity.

Many community detection methods have been proposed (He et al. 2016; Blondel et al. 2008; He et al. 2017; Jin et al. 2018; Ganji, Bailey, and Stuckey 2018; Jin et al. 2019), but most of them only rely on the lower-order structure at the level of individual nodes and edges, and ignore the higher-order structure in the network (Shi and Malik 2000; Newman 2006; Frey and Dueck 2007; Schaeffer 2007; Chakraborty et al. 2014). The common higher-order structures are small subgraphs in networks, also known as network motifs, which are crucial for understanding the fundamental structures and regulating the behavior of complex networks (Benson, Gleich, and Leskovec 2016). In order to capture the building blocks of the network, some community detection methods based on higher-order structure have been proposed in recent years (Arenas et al. 2008; Benson, Gleich, and Leskovec 2016; Tsourakakis, Pachocki, and Mitzenmacher 2017; Yin et al. 2017; Huang, Wang, and Chao 2018, 2019a,b). The basic idea of these higher-order methods is as follows: The first step is to construct a motif network by using motifs. If two nodes have involved in at least one common motif, there is a higher-order connection between them; otherwise, there is no higher-order connection. Then, lower-order methods are used on the motif adjacency matrix for community detection. However, this kind of approaches focuses on the motif network and ignores the the original lower-order topological structure of complex network. In the process of constructing motif network, some connected components and isolated nodes may be generated, which leads to the fragmentation issue of the motif network. At the same time, the construction principle of the motif network may cause two nodes connected in the original network not to be connected in the motif network, and making the nodes that might originally be in the same community belong to different communities, which violates the lower-order topological structure of the network, and causes the deviation of community detection results. Although a community detection method based on edge enhancement has proposed (Li et al. 2019a), the addition of edges may also break the original topological structure.

To make effective use of higher-order structure and lower-order structure for community detection, a Hybrid-order Stochastic Block Model (HSBM) is proposed. This method solves the problem of hybrid-order community detection from the perspective of the generative model. Based on the classical stochastic block model, the generation of lower-order structure and higher-order structure of the network is modeled uniformly. It is able to reveal the generation mechanism of the network from the perspective of community structure, and maintain the topological structure, higher-order structure, and statistical characteristics of the network. At the same time, a heuristic algorithm is proposed to op-

*Corresponding author

timize the objective function. Extensive experiments on six real-world datasets show that the proposed method is effective and advanced for community detection.

The main contributions of this paper are as follows:

- We propose a Hybrid-order Stochastic Block Model (HSBM) for the community detection, which leverages both higher-order and lower-order connectivity patterns.
- We for the first time propose to use the generative model to uniformly model the original network and motif network, avoiding the disseverance of higher-order and lower-order connectivity patterns and the fragmentation issues in the current community detection methods.
- Extensive experiments are conducted on six real-world networks to prove the effectiveness of the proposed method.

Preliminaries and Problem Statement

Before formally introducing the problem statement and the proposed approach, we briefly introduce some of the necessary background and notations.

The input is a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ means the node set of the graph, and $\mathcal{E} = \{e_1, e_2, ..., e_m\}$ denotes the edge set consisting of m edges. In this paper, we focus on undirected networks, which have been most extensively studied, and we allow the networks to include both multi-edges and self-edges without generality (Karrer and Newman 2010). $A \in \mathbb{N}^{n \times n}$ represents the adjacency matrix of the network, i.e. the lower-order connection of the original network. Specifically, A_{ij} represents the number of connections between node i and node j if i is not equal to j, and A_{ii} is equal to twice the number of self-edges of node i.

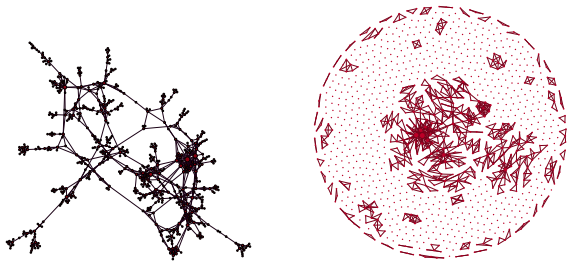
The lower-order structure above mainly focuses on the connectivity patterns at the level of individual nodes and edges, and ignores the higher-order connectivity patterns. Motif is a sub-network that occurs frequently in complex networks, and its number is significantly higher than that in randomized networks of the same degree distribution. As the building block of complex networks, it is crucial to understand the basic structure and to regulate the behavior of the networks (Benson, Gleich, and Leskovec 2016). Different motifs exist in different complex networks. For instance, the triangle motif has been widely found in social networks, and two-hop path motifs are common structures of air traffic networks. Formally, network motif can be expressed as:

$$\mathbf{M}_p^q = \{\mathcal{V}_\mathbf{M}, \mathcal{E}_\mathbf{M}\} \tag{1}$$

where $\mathcal{V}_\mathbf{M} \subseteq \mathcal{V}$ represents the node set consisting of p nodes, and $\mathcal{E}_\mathbf{M} \subseteq \mathcal{E}$ represents the edge set consisting of q edges in the motif $\mathbf{M}$. Given the original adjacency matrix A, we can define the motif adjacency matrix $M \in \mathbb{N}^{n \times n}$ based on different motif types, which is defined as:

$$M_{ij} = \text{number of motif instances containing nodes } i \text{ and } j \tag{2}$$

In this way, M is used to represent the motif-based higher-order connections of the complex network, where the edges represent the number of co-occurrence of nodes i and j under a specific motif type. Note that M_{ij} can be 0 even if $A_{ij} \neq 0$. This paper focuses on the triangular motif $\mathbf{M}_3^3$, but the proposed method can be well extended to other motifs.



(a) Original network (b) Motif network

Figure 1: Illustration of the differences between the original network and the motif network on the DBLP dataset.

From the above formula, we can see that in the process of constructing the motif network, the original lower-order topological structure is not well maintained. For illustration purpose, we illustrate the original network of the DBLP dataset and its corresponding motif network, as shown in Figure 1. From the figure, the structure of the original network and the motif network is significantly different. On the one hand, for the node-level, the lack of higher-order connectivity patterns between some nodes led to the following situation: two nodes are connected in the original network, while there are no edges between them in the motif network. If only the higher-order connectivity structure obtained by the above definition is used for the subsequent analysis, the original lower-order information will obviously be ignored, which will result in the deviation of community detection results. On the other hand, for the network-level, when we construct a motif network, the original connected network is divided into several connected components of different sizes and some isolated nodes, resulting in the fragmentation issue (Li et al. 2019a). In the process of partitioning nodes, these isolated nodes will not be supported by the original network, which makes the community labels of isolated nodes present randomness.

Although an edge enhancement approach has been proposed to solve the above issue (Li et al. 2019a), the added edges of the method may still destroy the original lower-order connectivity patterns. To our best knowledge, there is still a lack of methods to effectively utilize higher-order structure and lower-order structure for community detection. Since we are using both higher-order connectivity patterns and lower-order connectivity patterns, the proposed method should be named "Hybrid-order Stochastic Block Model (HSBM)". On the one hand, we maintain the topological structure of the original network so that the community assignments of nodes does not violate the original lower-order connectivity patterns. On the other hand, we utilize higher-order connectivity patterns, and construct the motif network to correct the deviation of community detection results caused by using only lower-order structure.

The Proposed Method

Motif Network Construction

We use the topological structure of the original network to construct a motif network, which is represented as:

$$\mathcal{G}^M = \{\mathcal{V}^M, \mathcal{E}^M\} \tag{3}$$

where $\mathcal{G}^M$ represents a motif network based on the triangle motif, and $\mathcal{V}^M$ represents the node set of the motif network containing n nodes, which is the same as the node set $\mathcal{V}$ in the original network, and $\mathcal{E}^M$ represents the edge set containing m' edges. In particular, we have $\mathcal{E}^M = \{\mathcal{E}_1^M, \mathcal{E}_2^M, ..., \mathcal{E}_{m'}^M\}$ with

$$\mathcal{E}_l^M = (i, j, M_{ij}), l = 1, 2, ..., m'. \tag{4}$$

where i and j represent the nodes at both ends of edges, and M_{ij} represents the number of edges between nodes i and j on the motif network as Eq. (2).

Hybrid-order Stochastic Block Model

In this section, we introduce the proposed Hybrid-order Stochastic Block Model (HSBM) in detail.

We first introduce the generation mechanism of the motif network. In the motif network, we assume that the number of edges generated between each pair of nodes follows an independent Poisson distribution. Let $g_i \in \{1, 2, ..., K\}$, $\forall i = 1, ..., n$ represent the community label of node i, where K is the number of communities. $X \in \mathbb{R}^{K\times K}$ represents the edge expectation matrix of nodes between communities in the motif network, where X_{rs} denotes the expected value of the motif adjacency matrix element M_{ij} for node i and j belonging to communities r and s respectively. To reflect the difference of nodes within the community in the motif network, the parameter $\xi \in [0, 1]^n$ is introduced to control the expected degrees of nodes. Therefore, in the motif network, the expected number of edges between nodes i and j belonging to communities r and s respectively is $\xi_i \xi_j X_{rs}$.

In this way, under the conditions of given parameters X, ξ and community labels g, the probability distribution of generating the motif network $\mathcal{G}^M$ can be obtained as follows:

$$P(M|\xi, X, g) = \prod_{i<j} \frac{(\xi_i \xi_j X_{g_i g_j})^{M_{ij}}}{M_{ij}!} \exp(-\xi_i \xi_j X_{g_i g_j}) \prod_i \frac{(\frac{1}{2}\xi_i^2 X_{g_i g_i})^{M_{ii}/2}}{(M_{ii}/2)!} \exp(-\frac{1}{2}\xi_i^2 X_{g_i g_i}) \tag{5}$$

where the expected number of self-edges at node i in community r is $1/2\xi_i^2 X_{rr}$. For each community r, there is a constraint about ξ as follows:

$$\sum_i \xi_i \delta_{g_i, r} = 1 \tag{6}$$

where δ is the Kronecker delta, which is equal to 1 when two subscripts are equal, and 0 otherwise. Then, the sum of ξ_i of all nodes in the community is 1, i.e., the value of ξ_i is equal to the probability that the connected node is i itself when an edge is connected to the community which node i belongs to in the motif network.

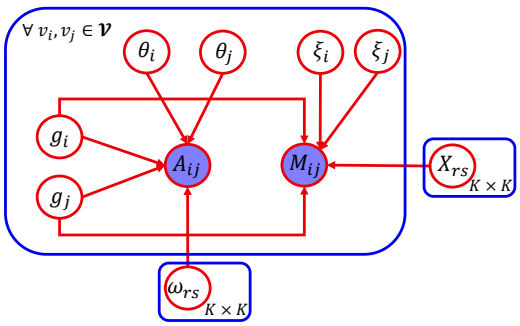


Figure 2: A graphical representation of HSBM

For the original adjacency matrix A_{ij}, the above process is also applicable. The generation probability of A can be obtained as follows:

$$P(A|\theta, \omega, g) = \prod_{i<j} \frac{(\theta_i \theta_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp(-\theta_i \theta_j \omega_{g_i g_j}) \prod_i \frac{(\frac{1}{2}\theta_i^2 \omega_{g_i g_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp(-\frac{1}{2}\theta_i^2 \omega_{g_i g_i}) \tag{7}$$

where $\omega \in \mathbb{R}^{K\times K}$ is the expected value of the original adjacency matrix elements. $\theta \in [0, 1]^n$ is the parameter of node differences in original networks, corresponding to ξ in the motif network, and its constraint is:

$$\sum_i \theta_i \delta_{g_i, r} = 1 \tag{8}$$

Figure 2 shows the generation process of the original adjacency matrix A and the motif adjacency matrix M, and reflects the conditional independent relationship between the observation variables A and M, and parameters g, θ, ξ, ω and X. It can be seen that the generation of the original adjacency matrix A is independent of other variables except for the community labels g and the expected value of edges ω, and the node difference parameter θ. The generation of motif adjacency matrix M only depends on the node labels g, the expected value of edges X, and the node difference parameter ξ in the motif network.

The proposed model generates both the original adjacency matrix A and the motif adjacency matrix M. In classical SBM, the generation of edges in the network is conditional independent. The motif network characterizes local motif structures, and the generation of M can alleviate the constraint of conditional independence in SBM. Although the model does not directly model the relationship between A and M, M is calculated by deterministic rules through A, and the community labels g are used to simultaneously model the generation of A and M, so that the two have mutually reinforcing effects. The model not only describes the original lower-order topological structure, but also realizes the influence of the higher-order connectivity pattern on the community structure.

Therefore, we can obtain the joint probability distribution of the observation variables A and M given the model parameters and community labels as follows:

$$\begin{aligned} P(A,M|\theta,&\omega,g,\xi,X) \\ &=P(A|\theta,\omega,g)P(M|\xi,X,g) \\ &=\frac{1}{\prod_{i<j}A_{ij}!\prod_i 2^{A_{ii}/2}(A_{ii}/2)!} \\ &\quad\frac{1}{\prod_{i<j}M_{ij}!\prod_i 2^{M_{ii}/2}(M_{ii}/2)!} \\ &\quad\prod_i \theta_i^{k_i}\prod_{rs}\omega_{rs}^{m_{rs}/2}exp\{-\frac{1}{2}\omega_{rs}\} \\ &\quad\prod_i \xi_i^{k'_i}\prod_{rs}X_{rs}^{m'_{rs}/2}exp\{-\frac{1}{2}X_{rs}\} \end{aligned} \tag{9}$$

where k_i and k'_i are the degree of node i in the original network and the motif network respectively. m_{rs} and m'_{rs} are the total number of edges between communities r and s in the original network and the motif network, respectively, i.e.

$$m_{rs}=\sum_{ij}A_{ij}\delta_{g_i,r}\delta_{g_j,s},\quad m'_{rs}=\sum_{ij}M_{ij}\delta_{g_i,r}\delta_{g_j,s} \tag{10}$$

Notice that when r is equal to s, m_{rs} and m'_{rs} are twice the total number of edges within the community in the corresponding network above.

The goal is to maximize the probability in Eq. (9) with respect to the unknown parameters ω, θ, X, ξ and the community labels g. It can be calculated more easily by maximizing the logarithm of the probability. The logarithmic of Eq. (9) that neglects constants is as follows:

$$\begin{aligned} &\log P(A,M|\theta,\omega,g,\xi,X) \\ &=2\sum_i k_i\log\theta_i+\sum_{rs}(m_{rs}\log\omega_{rs}-\omega_{rs}) \\ &+2\sum_i k'_i\log\xi_i+\sum_{rs}(m'_{rs}\log X_{rs}-X_{rs}) \end{aligned} \tag{11}$$

Under the constraints of Eq. (6) and Eq. (8), the model parameters obtained by maximizing logarithmic likelihood are as follows:

$$\hat{\theta}_i=\frac{2k_i}{\sum_i 2k_i\delta_{g_i,r}}=\frac{k_i}{\mathcal{K}_{g_i}},\quad \hat{\xi}_i=\frac{k'_i}{\mathcal{K}'_{g_i}} \tag{12}$$

$$\hat{\omega}_{rs}=m_{rs},\quad \hat{X}_{rs}=m'_{rs} \tag{13}$$

where $\mathcal{K}_r$ and $\mathcal{K}'_r$ are the degrees of all nodes in community r in the original network and motif network respectively, i.e.

$$\mathcal{K}_r=\sum_s m_{rs}=\sum_i k_i\delta_{g_i,r},\quad \mathcal{K}'_r=\sum_i k'_i\delta_{g_i,r} \tag{14}$$

And they are also equal to the total number of edges generated by nodes in community r in the corresponding networks.

By substitute Eq. (12)- Eq. (13) into $\log P(A,M)$, we get:

$$\begin{aligned} &\log P(A,M|\theta,\omega,g,\xi,X) \\ &=2\sum_i k_i\log\frac{k_i}{\mathcal{K}_{gi}}+\sum_{rs}m_{rs}\log m_{rs}-2m \\ &+2\sum_i k'_i\log\frac{k'_i}{\mathcal{K'}_{gi}}+\sum_{rs}m'_{rs}\log m'_{rs}-2m' \end{aligned} \tag{15}$$

where m and m' are the total number of edges in the original network and motif network, respectively. Further simplifying Eq. (15), the objective function can be obtained as:

$$\mathcal{L}(A,M|g)=\sum_{rs}m_{rs}\log\frac{m_{rs}}{\mathcal{K}_r\mathcal{K}_s}+\sum_{rs}m'_{rs}\log\frac{m'_{rs}}{\mathcal{K}'_r\mathcal{K}'_s} \tag{16}$$

Since the calculation of Eq. (16) needs to consider the situation of the whole network each time, the calculation is relatively complicated. In order to improve the efficiency of the algorithm, we calculate the change of Log-likelihood when a node changes its community. When node i is transferred from community r to community s, the change of log likelihood can be written as:

$$\begin{aligned} &\Delta\mathcal{L} \\ &=\sum_{t\neq r,s}[a(m_{rt}-k_{it})-a(m_{rt})+a(m_{st}+k_{it})-a(m_{st})] \\ &+a(m_{rs}+k_{ir}-k_{is})-a(m_{rs})+b[m_{rr}-2(k_{ir}+u_i)] \\ &-b(m_{rr})+b[m_{ss}+2(k_{is}+u_i)]-b(m_{ss}) \\ &-a(\mathcal{K}_r-k_i)+a(\mathcal{K}_r)-a(\mathcal{K}_s+k_i)+a(\mathcal{K}_s) \\ &+\sum_{t\neq r,s}[a(m'_{rt}-k'_{it})-a(m'_{rt})+a(m'_{st}+k'_{it})-a(m'_{st})] \\ &+a(m'_{rs}+k'_{ir}-k'_{is})-a(m'_{rs})+b[m'_{rr}-2(k'_{ir}+u'_i)] \\ &-b(m'_{rr})+b[m'_{ss}+2(k'_{is}+u'_i)]-b(m'_{ss}) \\ &-a(\mathcal{K}'_r-k'_i)+a(\mathcal{K}'_r)-a(\mathcal{K}'_s+k'_i)+a(\mathcal{K}'_s) \end{aligned} \tag{17}$$

Here, we define $a(x)=2x\log x$, $b(x)=x\log x$, and $a(0)=0$, $b(0)=0$. k_{it} is the number of edges between node i and the nodes that belong to community t in the original network, and u_i is the number of self-edges of node i. All symbols with superscript such as k'_{it} and u'_i represent their equivalents in the motif network.

Algorithm Summary and Analysis

According to the community labels of nodes $g_i,\forall i=1,...,n$, the community structure of the network can be obtained, which is denoted by $\mathcal{C}=\{\mathcal{C}_1,...,\mathcal{C}_K\}$, where K is the number of communities in the network.

$$\mathcal{C}_r=\{v_i|v_i\in\mathcal{V},s.t.\ g_i=r\},\ \forall r=1,..,K. \tag{18}$$

For clarity, we summarize the main procedure of the proposed HSBM in Algorithm 1.

We analyze the complexity of the proposed HSBM method as follows. In general, the complexity of this algorithm mainly depends on two parts: calculating motif adjacency matrix and partition nodes to communities in the network. For triangle motif, the worst-case computational complexity of constructing motif adjacency matrix is $O(m^{1.5})$,

Algorithm 1 Hybrid-order Stochastic Block Model

Require: Adjacency matrix of network A, number of communities K, stop criterion ε.

1: Construct motif adjacency matrix M from A via Eq. (2).
2: Obtain the initial community labels g by random initialization.
3: Compute objective score $\mathcal{L}^{new}$ via Eq. (16).
4: **repeat**
5: $\quad\mathcal{L}^{old} = \mathcal{L}^{new}$.
6: $\quad$**for** every node i **do**
7: $\quad\quad g_i \leftarrow \arg\max_s \Delta\mathcal{L}(s)$ via Eq. (17).
8: $\quad$**end for**
9: $\quad g \leftarrow \arg\max_g \mathcal{L}$
10: **until** $|\mathcal{L}^{new} - \mathcal{L}^{old}| < \varepsilon$
11: Convert community labels $g_i, \forall i = 1, ..., n$ to community structure $\{\mathcal{C}_1, ..., \mathcal{C}_K\}$ of the network via Eq. (18).

Ensure: Community labels g.

where m is the number of edges in the original network. In practice, the computation is faster, and the computational complexity is $O(m^{1.2})$ (Benson, Gleich, and Leskovec 2016). The complexity of computing $\Delta\mathcal{L}$ is $O(K + \langle k\rangle + \langle k'\rangle)$, where $\langle k\rangle$ and $\langle k'\rangle$ are the average degree of nodes in the original network and motif network respectively. Given node i and the original community r, finding the community s that maximizes $\Delta\mathcal{L}$ requires $O[K(K + \langle k\rangle + \langle k'\rangle)]$. The time required to assign each node in the network to the communities is $O[N(K(K + \langle k\rangle + \langle k'\rangle))]$. Therefore, the overall complexity of Algorithm 1 is $O[m^{1.2} + N(K(K + \langle k\rangle + \langle k'\rangle))]$.

Experiments

Experimental Setting

Datasets Six widely used real-world datasets are adopted to test the effectiveness of the proposed method.

- **KareteClub**[1]: A social network about karate clubs composed of 34 nodes and 78 edges, and the nodes are divided into 2 communities.
- **Polbooks**[1]: A book network about US politics composed of 105 nodes and 441 edges, and the nodes are divided into 3 communities.
- **Polblogs**[1]: A network of hyperlinks between weblogs on US politics composed of 1490 nodes and 19090 edges, and the nodes are divided into 2 communities.
- **Dolphins**[1]: A dolphin social network of frequent associations composed of 62 nodes and 159 edges, and the nodes are divided into 2 communities.
- **Football**[1]: A social network about the American college football league composed of 115 nodes and 616 edges, and the nodes are divided into 12 communities.
- **DBLP**[2]: A paper cooperative network, which is a subset of DBLP dataset, containing 1163 nodes and 1392 edges, and the nodes are divided into 3 communities.

[1]http://www-personal.umich.edu/ mejn/netdata/
[2]http://snap.stanford.edu/data/

Baselines Seven lower-order community detection methods and five higher-order community detection methods are used as baseline. The lower-order methods include Stochastic block model (**SBM**), **Louvain**, Nonnegative Matrix Factorization (**NMF**), Spectral Clustering (**SC**), Spectral clustering based on normalized cut (**Ncut**), Affinity propagation (**AP**), and **FastNewman** (Blondel et al. 2008; Wang et al. 2011; Ng, Jordan, and Weiss 2001; Shi and Malik 2000; Frey and Dueck 2007; M. et al. 2004). The higher-order methods as follows:

- **EdMot**: An edge enhancement method for motif-aware community detection, which solves the hypergraph fragmentation issue by adding new edges to construct motif hypergraphs (Li et al. 2019a). The new adjacency matrix constructed by EdMot are adopted as the input of the lower-order methods, resulting in the new higher-order methods, namely **EdMot-Louvain**, **EdMot-NMF** and **EdMot-SC**, respectively.
- **MWLP**: A method based on label propagation, which integrates higher-order structure features, designs a unified re-weighted network of higher-order structure and lower-order structure, and updates node communities based on label propagation process (Li et al. 2019b).
- **Motif-Cond**: A community detection method based on higher-order connectivity patterns. By extending the spectral clustering method based on eigenvalues and eigenvectors, the higher-order structure is adopted to obtain the optimal partition of the network (Benson, Gleich, and Leskovec 2016).

Performance Metrics Three commonly used performance metrics for community detection are adopted to evaluate the effectiveness of the proposed method (Chakraborty et al. 2014), which are normalized mutual information (NMI), F1-Score and Modularity respectively. NMI evaluates the similarity between the predicted communities and the ground-truth communities from the perspective of information theory. F1-score considers the harmonic values of Precision and Recall comprehensively. Modularity is used to measure the closeness of community structure. The first two require the ground-truth, and the value range is $[0, 1]$. The last one does not require, and the value range is $[-0.5, 1)$.

Comparison Results

In this section, we compare HSBM with twelve methods, including the lower-order community detection methods and higher-order community detection methods.

Figure 3 shows the comparison results in terms of NMI, F1-Score and Modularity, respectively. As can be seen from the figure, higher-order methods generally obtain better performance than lower-order methods, which reflects the effectiveness and necessity of combining motif to conduct community detection. By adding information about higher-order connectivity patterns, blocks of networks are more easily captured.

HSBM has advantages in terms of NMI and F1-Score on polblogs and DBLP networks. There is a mass of motif

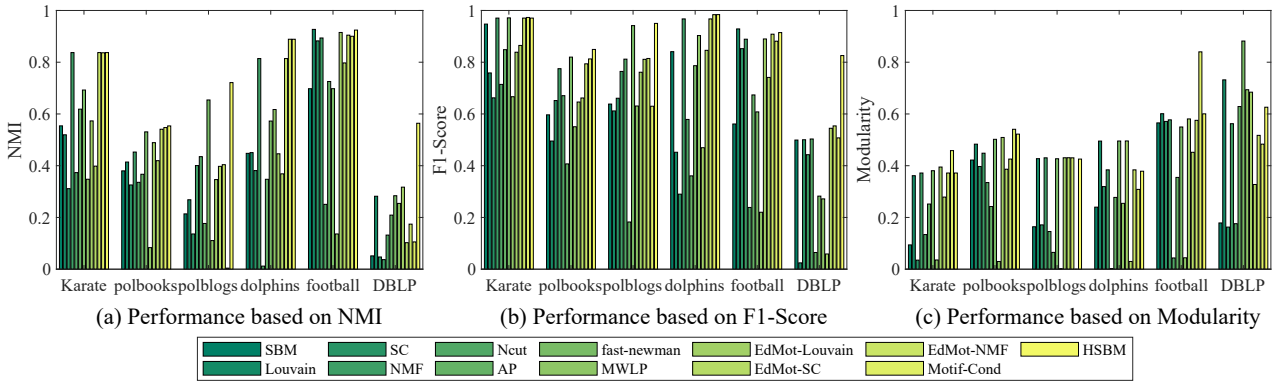


Figure 3: Comparison results with the twelve community detection methods on the six real-world networks.

structures in these two networks, and the fusion of higher-order connectivity patterns in HSBM improves the performance of community detection and achieves better results than lower-order methods. Due to the obvious fragmentation issue of these two networks, the original network generated by HSBM maintained the lower-order structure, alleviated the fragmentation issue of motif networks in higher-order methods, and corrected the deviation of community detection results caused by relying only on the higher-order connectivity patterns. In the polbooks network that does not suffer the fragmentation issue, the proposed HSBM method still performs well. In the football network, HSBM and the higher-order methods have similar results. Due to the group stage principle of the football league, the football dataset shows obvious community structure in the motif network, so whether the original topological structure is adopted has little impact on the final results of community detection.

In general, the proposed HSBM method has better performance than baselines. Because the lower-order methods do not fuse motif, they cannot reflect the higher-order structure. Although the motif based higher-order methods derive some new connections, in general, they only use motif network constructed by higher-order connectivity patterns, without considering the original topological structure.

Case Study

In this subsection, case study is conducted on the polblogs dataset to verify the validity of the proposed HSBM method. This dataset comes from web blogs around the time of the 2004 US presidential election, about the political leanings of blogs and their online connections. On the dataset, each blog is labeled Liberal or Conservative, and we selected the largest connection component with 1,222 vertices for demonstration and analysis.

Experiments show that the proposed HSBM method outputs a better partition of the communities, which is closer to ground-truth. Two typical nodes are selected to illustrate the effectiveness of the proposed method. To facilitate understanding and presentation, the self-networks of corresponding nodes are captured in Figure 4 and Figure 5, showing the community labels given by ground-truth, HSBM, and baseline methods, as well as the corresponding motif networks.

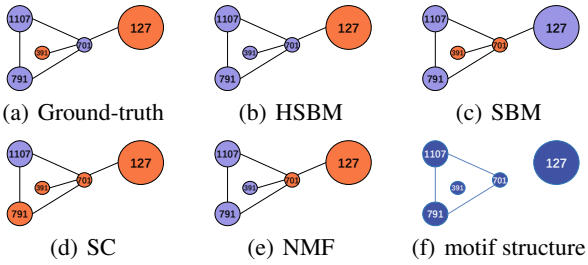
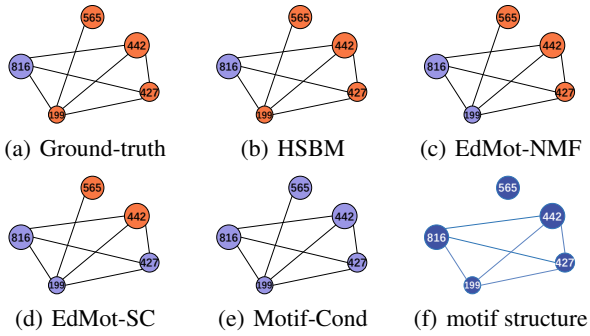


Figure 4: Visualization of community detection on polblogs for node 701. The size of each node is proportional to its degree. Different colors represent different community labels.

Compared with the lower-order methods, HSBM adds the higher-order connectivity patterns ignored in the lower-order methods to improve the effectiveness of community detection. As shown in Figure 4, node 701 was incorrectly classified as Liberal Party in all the lower-order methods, and HSBM correctly identified it as the Conservative Party. By analyzing its original topological structure, it is found that a hub node of the Liberal Party 127 connected to node 701 has a significant influence on its results. HSBM incorporates the higher-order connectivity patterns, in the constructed motif network, node 701 had no higher-order connection with node 127, which alleviated the excessive influence of node 127 on the community label of node 701 and corrected the deviation of community detection results. For node 127, which all other methods gave the correct labels, SBM assigned it to the Conservative Party because the method tends to divide nodes of similar degree into the same community.

Compared with the higher-order method, HSBM corrects the deviation of community detection results caused by excessive reliance on higher-order connectivity patterns and violation of the lower-order structure in the higher-order methods. As shown in Figure 5, node 199 was incorrectly classified as Conservative Party in all the higher-order methods, but HSBM correctly identified it as the Liberal party. In the constructed motif network, compared with the original network, the connection between the node 199 and the Liberal node was reduced, but the connection with the Con-



(a) Ground-truth (b) HSBM (c) EdMot-NMF

(d) EdMot-SC (e) Motif-Cond (f) motif structure

Figure 5: Visualization of community detection on polblogs for node 199.

servative node was not affected. In the higher-order methods that relied only on motif network, it was overly influenced by conservative nodes, leading to bias in the community detection results. While HSBM uses the higher-order connectivity patterns, it does not discard the lower-order connectivity patterns, which maintains the original topological structure.

In general, HSBM corrects some of the overzealous tendencies in higher-order or lower-order methods. The result confirms the effectiveness and necessity of the integration of the higher-order and lower-order connectivity patterns.

Related Work

Generative Model for Community Detection

The generative model has been widely studied and developed in the field of community detection due to its excellent interpretability and theoretical basis. The main idea is to build a complex network generative model with community structure, and make statistical inferences on the parameters in the network to get the results of community detection.

In the classical stochastic block model, the nodes in the network are randomly divided into different communities, and then the nodes in the communities are regarded as undifferentiated to generate edges. The degree-corrected block model is proposed to correct the deviations of the community detection results of the stochastic block model, and the degree heterogeneity of nodes is captured by introducing the parameters that control the expected degrees of nodes (Karrer and Newman 2010). PLD-SBM introduces latent variables to represent the power-law distribution of node degree in the real world, so that the community detection results are closer to the ground-truth (Qiao et al. 2019).

In addition, some efforts have been made in developing stochastic block models with mixed membership (Airoldi et al. 2008), stochastic block models for model selection (Chen, Zhang, and Xiong 2016), and latent space model based methods (Sewell and Chen 2016). However, the above methods only take advantage of the original lower-order structure and ignore the higher-order connectivity patterns.

Motif-based Higher-order Community Detection

There are abundant higher-order organizational structures in real networks, and higher-order connectivity patterns play a vital role in understanding and controlling the behavior of complex systems. Motif refers to the most common higher-order structure in networks, which is defined as the connectivity pattern in networks with a significantly larger occurrence number than that in randomized networks preserving the same degree of nodes. The motif is widely used to reveal the generation mechanism of complex networks.

Different from the lower-order community detection methods, the higher-order methods utilize motif in the network to obtain communities. A generalized clustering network framework based on the higher-order connectivity pattern was proposed in (Benson, Gleich, and Leskovec 2016). By integrating motif, the spectral clustering method is extended to obtain community partition so as to reveal higher-order organizations in the network. A general motif-based framework was proposed in (Arenas et al. 2008), which conducted community detection by using motif instead of edges as the basic unit in networks and expanding the modularity proposed in (Newman 2006). The EdMot method designed an edge enhancement strategy (Li et al. 2019a), which derived new edges in motif-based hypergraphs to solve the hypergraph fragmentation issue caused by the use of only higher-order connectivity patterns. The MuMod method adopted the higher-order and lower-order connectivity patterns to construct the micro-unit connection network, and through the micro-unit module model, and the overlapping community structure of the network is obtained through a micro-unit modularity model (Huang, Chao, and Xie 2020). In addition, some methods using motifs have been extended in the direction of local higher-order graph partitioning (Yin et al. 2017).

Most of the existing higher-order methods are based on motif to construct new motif network, and then utilize the lower-order community detection methods to detect communities on the motif network. Although higher-order building blocks are captured, the lower-order connectivity patterns are ignored or even violated. As a result, this paper proposes Hybrid-order Stochastic Block Model, where both of the original lower-order topological structures and the higher-order structures based on motif are modeled under a unified framework, so as to simultaneously take into account the higher-order and lower-order connectivity patterns.

Conclusion

In this paper, we propose a Hybrid-order Stochastic Block Model (HSBM) for community detection. Different from the existing higher-order community detection methods, we for the first time propose to use the generative model to uniformly model the original network and motif network. In the proposed model, both the higher-order and lower-order connectivity patterns are utilized to influence the community labels of nodes simultaneously. Extensive experiments have shown the effectiveness of the proposed method.

Acknowledgments

This project was supported by NSFC (61876193, 61902278), and Guangdong Natural Science Funds for Distinguished Young Scholar (2016A030306014).

References

Airoldi, E. M.; Blei, D. M.; Fienberg, S. E.; and Xing, E. C. 2008. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research Jmlr* 9(5): 1981–2014.

Arenas, A.; Fernandez, A.; Fortunato, S.; and Gomez, S. 2008. Motif-based communities in complex networks. *Journal of Physics A: Mathematical and Theoretical* 41(22): 224001.

Benson, A. R.; Gleich, D. F.; and Leskovec, J. 2016. Higher-order organization of complex networks. *Science* 353(6295): 163–166.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008(10): P10008.

Chakraborty, T.; Srinivasan, S.; Ganguly, N.; Mukherjee, A.; and Bhowmick, S. 2014. On the permanence of vertices in network communities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1396–1405.

Chen, G.; Zhang, H.; and Xiong, C. 2016. Maximum margin Dirichlet process mixtures for clustering. In *Thirtieth AAAI Conference on Artificial Intelligence*, 1491–1497.

Frey, B. J.; and Dueck, D. 2007. Clustering by passing messages between data points. *science* 315(5814): 972–976.

Ganji, M.; Bailey, J.; and Stuckey, P. J. 2018. Lagrangian constrained community detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2983–2990.

He, D.; Feng, Z.; Jin, D.; Wang, X.; and Zhang, W. 2017. Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In *Thirty-First AAAI Conference on Artificial Intelligence*, 116–124.

He, L.; Lu, C.-T.; Ma, J.; Cao, J.; Shen, L.; and Yu, P. S. 2016. Joint community and structural hole spanner detection via harmonic modularity. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 875–884.

Huang, L.; Chao, H.-Y.; and Xie, G. 2020. MuMod: A Micro-Unit Connection Approach for Hybrid-Order Community Detection. In *AAAI*, 107–114.

Huang, L.; Wang, C.-D.; and Chao, H.-Y. 2018. A harmonic motif modularity approach for multi-layer network community detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1043–1048.

Huang, L.; Wang, C.-D.; and Chao, H.-Y. 2019a. Higher-order multi-layer community detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9945–9946.

Huang, L.; Wang, C.-D.; and Chao, H.-Y. 2019b. HM-Modularity: A Harmonic Motif Modularity Approach for Multi-layer Network Community Detection. *IEEE Transactions on Knowledge and Data Engineering* .

Jin, D.; Wang, X.; He, R.; He, D.; Dang, J.; and Zhang, W. 2018. Robust detection of link communities in large social networks by exploiting link semantics. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 314–321.

Jin, D.; You, X.; Li, W.; He, D.; Cui, P.; Fogelman-Soulié, F.; and Chakraborty, T. 2019. Incorporating network embedding into markov random field for better community detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 160–167.

Karrer, B.; and Newman, M. E. J. 2010. Stochastic blockmodels and community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 83(2): 016107.

Li, P.-Z.; Huang, L.; Wang, C.-D.; and Lai, J.-H. 2019a. EdMot: An Edge Enhancement Approach for Motif-aware Community Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 479–487.

Li, P. Z.; Huang, L.; Wang, C. D.; Lai, J. H.; and Huang, D. 2019b. Community Detection by Motif-Aware Label Propagation. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14(2): 22:1–22:19.

M.; E.; J.; and Newman. 2004. Fast algorithm for detecting community structure in networks. *Physical Review E* 69(6 Pt 2): 066133.

Newman, M. E. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103(23): 8577–8582.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2001. On Spectral Clustering: Analysis and an Algorithm. In *Adv. Neural Inf. Proc. Syst. 14*.

Qiao; Maoying; Jun; Bian; Wei; Qiang; Tao; and Dacheng. 2019. Adapting Stochastic Block Models to Power-Law Degree Distributions. *IEEE Transactions on Cybernetics* .

Schaeffer, S. E. 2007. Graph clustering. *Computer science review* 1(1): 27–64.

Sewell, D. K.; and Chen, Y. 2016. Latent space models for dynamic networks with weighted edges. *Social Networks* 44: 105–116.

Shi, J.; and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22(8): 888–905.

Tsourakakis, C. E.; Pachocki, J.; and Mitzenmacher, M. 2017. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*, 1451–1460.

Wang, F.; Li, T.; Wang, X.; Zhu, S.; and Ding, C. 2011. Community discovery using nonnegative matrix factorization. *Data Mining & Knowledge Discovery* 22(3): 493–521.

Yin, H.; Benson, A. R.; Leskovec, J.; and Gleich, D. F. 2017. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 555–564.