Community Detection by Motif-Aware Label Propagation

PEI-ZHEN LI, LING HUANG, CHANG-DONG WANG, and JIAN-HUANG LAI,
Sun Yat-sen University
DONG HUANG, South China Agricultural University

Community detection (or graph clustering) is crucial for unraveling the structural properties of complex networks. As an important technique in community detection, label propagation has shown the advantage of finding a good community structure with nearly linear time complexity. However, despite the progress that has been made, there are still several important issues that have not been properly addressed. First, the label propagation typically proceeds over the lower order structure of the network and only the direct one-hop connections between nodes are taken into consideration. Unfortunately, the higher order structure that may encode design principle of the network and be crucial for community detection is neglected under this regime. Second, the stability of the identified community structure may also be seriously affected by the inherent randomness in the label propagation process. To tackle the above issues, this article proposes a *Motif-Aware Weighted Label Propagation* method for community detection. We focus on triangles within the network, but our technique extends to other kinds of motifs as well. Specifically, the motif-based higher order structure mining is conducted to capture structural characteristics of the network. First, the motif of interest (locally meaningful pattern) is identified, and then, the motif-based hypergraph can be constructed to encode the higher order connections. To further utilize the structural information of the network, a re-weighted network is designed, which unifies both the higher order structure and the original lower order structure. Accordingly, a novel voting strategy termed *NaS* (considering both Number and Strength of connections) is proposed to update node labels during the label propagation process. In this way, the random label selection can be effectively eliminated, yielding more stable community structures. Experimental results on multiple real-world datasets have shown the superiority of the proposed method.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Community detection, higher order structure, motifs, label propagation

ACM Reference format:
Pei-Zhen Li, Ling Huang, Chang-Dong Wang, Jian-Huang Lai, and Dong Huang. 2020. Community Detection by Motif-Aware Label Propagation. *ACM Trans. Knowl. Discov. Data* 14, 2, Article 22 (February 2020), 19 pages.
https://doi.org/10.1145/3378537

This project was supported by NSFC (61876193 and 61976097), Guangdong Natural Science Funds for Distinguished Young Scholar (2016A030306014), Tip-top Scientific and Technical Innovative Youth Talents of Guangdong special support program (2016TQ03X542), and the Fundamental Research Funds for the Central Universities (19lgjc10).
Authors' addresses: P.-Z. Li, L. Huang, and C.-D. Wang, School of Data and Computer Science, Sun Yat-sen University, Guangdong Province Key Laboratory of Computational Science, Guangzhou Higher Education Mage Center, Panyu District, Guangzhou, P.R. China, 510006; emails: sysuLiPeizhen@163.com, huanglinghl@hotmail.com, changdongwang@hotmail.com; J.-H. Lai, School of Data and Computer Science, Sun Yat-sen University, Guangdong Key Laboratory of Information Security Technology, Guangzhou Higher Education Mage Center, Panyu District, Guangzhou, P.R. China, 510006; email: stsljh@mail.sysu.edu.cn; D. Huang, College of Mathematics and Informatics, South China Agricultural University, No. 483, Wushan Road, Guangzhou, P.R. China, 510642; email: huangdonghere@gmail.com.

1556-4681/2020/02-ART22 $15.00
https://doi.org/10.1145/3378537

1 INTRODUCTION

Many complex systems can naturally be represented by graphs or networks, where the nodes stand for elementary units of the system and the edges represent their relations. The examples include, but are not limited to, networks of webpages, scientific citation networks, social networks, and so on. Many real-world networks are found to have *community structures*, i.e., groups of nodes with relatively denser intra-group connections but sparser inter-group connections [10, 16, 46, 47, 49, 60]. Detecting communities (also called clusters or modules) in real-world networks have attracted significant attention [2, 4–6, 8, 9, 43] and many community detection methods have been proposed in the past few decades [1, 29, 33, 34, 40, 45, 48]. One of the important categories in community detection is based on the optimization of some fitness measures, e.g., modularity [29], Ncut [40], and so on. Another important community detection category is based on the non-negative matrix factorization (NMF) [48]. Despite significant success, the fitness measure based methods and the NMF based methods are often restricted by their high computational complexity in real applications. To deal with large-scale networks, the label propagation-based algorithms such as SLPA [57] and RAK [33] have proved to be powerful techniques for community detection due to the simplicity and nearly linear time complexity [33]. However, these methods mostly focus on the lower order structure of the network (at the level of the direct one-hop connections), while the connections revealed by higher order features (e.g., motifs) are often neglected. Although some motif-based community detection methods have been proposed [1, 13, 45], yet they generally suffer from high computational complexity when confronted with large-scale networks. It remains an open problem how to effectively and efficiently incorporate both lower order and higher order structural information in a unified community detection framework. In this article, we propose a *Motif-Aware Weighted Label Propagation* (*MWLP*) method for community detection, where the motif-based higher order features of the network are explored and fully utilized. Specifically, a re-weighted network is designed, which unifies both the higher order structure and the original lower order structure. The motif-based intimacy is integrated into lower order connection and can be reflected from the weight of the corresponding edge. Upon obtaining the re-weighted network, the community detection task can be performed through label propagation with the proposed voting strategy, i.e., *NaS* (considering both Number and Strength of connections). This paradigm allows for effective and efficient community detection as the higher order connections can provide more structural information of the network besides the lower order connections. Besides, the label propagation scheme endows the proposed *MWLP* with efficiency and thus makes it computationally feasible for large-scale networks. Note that we focus on triangles within the network, but our technique extends to other kinds of motifs as well. Extensive experiments on seven real-world datasets demonstrate that the proposed method achieves a better community detection performance over several baseline methods.

The rest of this article is organized as follows. In Section 2, we briefly review the background and make the problem statement. Section 3 details the proposed method, including motif-based higher order structure mining and motif-aware weighted label propagation. The experimental results are reported in Section 4. We draw the conclusion in Section 5.

2 BACKGROUND AND PROBLEM STATEMENT

2.1 Motifs in Community Detection

Motifs are representative higher order features of the network, which are defined as recurring, significant patterns of interconnections (compared with the corresponding random networks) or simply the building blocks of the network [3, 14, 15, 20, 21, 27]. Such higher order connectivity patterns are crucial for uncovering the organization of complex networks and provide insights in other

network research areas [58]. Thus, motif-based characteristics are promising for community detection and have been gaining increasing attention in community detection [1, 3, 45, 59]. A general motif modularity based framework was introduced in [1], which aims to find the modular structure of the network by exploiting the motif modularity derived from the original modularity [30]. A motif-based heuristic method termed *Triangle Connected Component Clustering* (TECTONIC) was proposed by generalizing the notion of conductance to triangle conductance [45]. Motif conductance was also defined in [3] to find the higher order organization of complex networks. In addition, great effort has been made in leveraging motifs in local higher order graph clustering [59], where the *Motif-based Approximate Personalized PageRank* (MAPPR) algorithm was proposed to quickly find a cluster containing a given seed node with minimal motif conductance. However, these methods are generally time consuming when dealing with large-scale networks as they typically rely on the optimization of certain fitness measure even without considering the computational cost on motif identification and the motif-based hypergraph construction.

2.2 Label Propagation

Label propagation-based community detection methods have been shown to perform efficiently for its simplicity [33, 55]. One typical label propagation algorithm is RAK, which identifies the community structures using the network structure alone as its guide [33]. The main advantage of RAK is that it has nearly linear time computational complexity, i.e., $O(\tau m)$, where τ is the number of propagation iterations and m is the number of edges in the network. RAK assumes that every node in the network carries a label denoting the community to which it belongs. For a specific node x, it will update its label according to its neighbors, and join the community that most of its neighbors belong to. If there are more than one communities with the maximum number of neighbors, then one of them will be randomly selected, which accounts for the uncertainty and randomness that root in the method. That is, different community structures can be obtained in different runs over the same network. More in detail, every node is initialized with unique labels, and as the labels propagate through the network, densely connected groups of nodes quickly reach a consensus on a unique label. Figure 1 illustrates this process intuitively. The label updating can be performed iteratively in synchronous or asynchronous schemes. In the synchronous updating scheme, node x in the tth iteration updates its label according to the labels of its neighbors in the $(t-1)$th iteration. In other words, $C_x(t) = f(C_{N_x^{(1)}}(t-1), \ldots, C_{N_x^{(\kappa)}}(t-1))$, where $N_x^{(i)}, \forall i \in \{1, \ldots, \kappa\}$ is the ith neighbor of node x, and κ denotes the number of neighbors. However, the oscillations of labels would happen if the subgraphs in the network are bi-partite or nearly bi-partite, as shown in Figure 2. Hence, we adopt the asynchronous updating scheme:

$$C_x(t) = f\left(C_{N_x^{(1)}}(t), \ldots, C_{N_x^{(l)}}(t), C_{N_x^{(l+1)}}(t-1), \ldots, C_{N_x^{(\kappa)}}(t-1)\right), \tag{1}$$

where $N_x^{(1)} \ldots N_x^{(l)}$ are the neighbors whose labels have already been updated in the current iteration and $N_x^{(l+1)} \ldots N_x^{(\kappa)}$ are the neighbors whose labels have not been updated yet. The iterative updating process will not stop until every node in the network does not change its label. In practice, it has been proved that 95% of nodes can be identified correctly after five iterations [33].

2.3 Issues and Challenges

Despite the simplicity and efficiency of the label propagation-based methods, we argue that there still are some issues that have not been well addressed. For instance, only lower order structure of the network (at the level of individual nodes and edges) is explored and utilized during the label propagation process while the higher order structure (at the level of small subgraphs, e.g., motifs) is not utilized. Actually, such higher order characteristics of the network can be quite

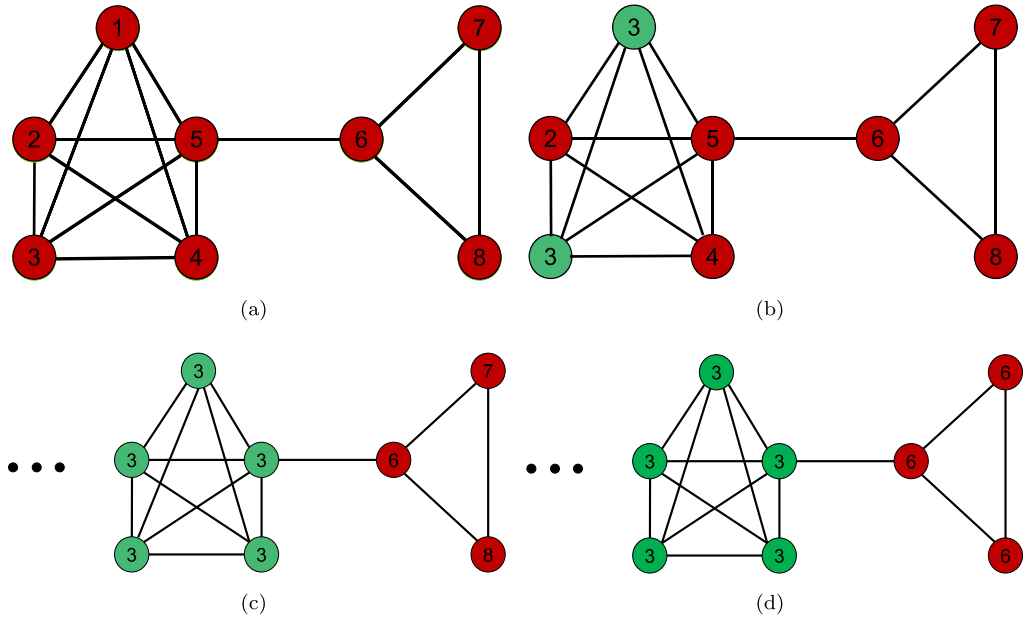


Fig. 1. Illustration of the label propagation process: two densely connected groups of nodes reach consensus on two unique labels respectively (highest possible in this case although the node is randomly picked and updated. The numbers denote node labels).

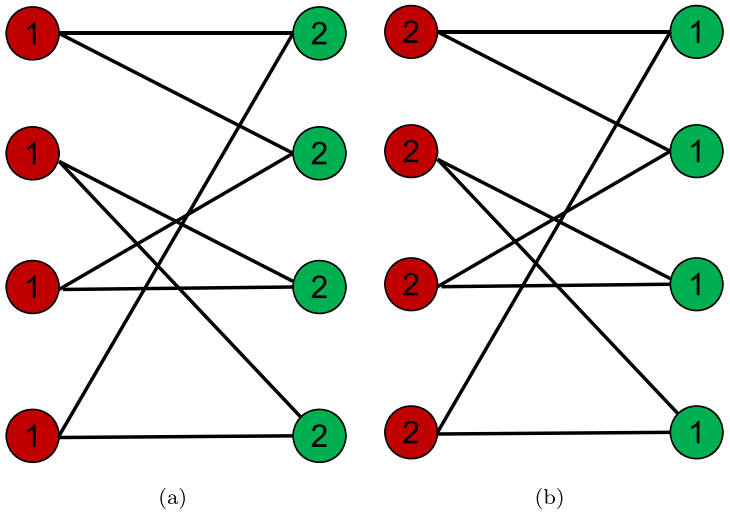


Fig. 2. An example of the label oscillation when the synchronous updating scheme is applied (the numbers denote node labels).

important when it comes to graph clustering. For one thing, the intimacy of nodes within a smaller neighboring area (specified by the motif of interest) is identified, which can distinguish different kinds of networks for their intrinsic structures; for another, noisy edges or some edges that are established due to occasional connections would be removed or attached with small weights, by which we can discriminate the strength or the importance of different connections. It is worth

noting that both two aforementioned aspects are crucial for a good graph partitioning. For clarity, we summarize the tricky issues that root in the basic label propagation-based methods as follows:

(1) In the label propagation process, each node adopts the label that is held by most of its neighbors, which means that only one-hop neighborhood structure is taken into consideration, while further connections between nodes are ignored, giving rise to the loss of multi-scale topological information of the network [25].
(2) Basic label propagation-based methods treat every connection between nodes equally, that is, the edges are unweighted. Since label propagation considers only the number of edges (without taking into consideration the importance of edges), different relations cannot be distinguished in this way.
(3) The stability of the basic label propagation-based methods is also an issue since the randomness is introduced in this algorithm. Specifically, different community structures could be discovered in different runs over the same network.

Although the weighted label propagation has been explored [44, 54] (corresponding to the second issue) and there have been some label propagation-based methods tailored to tackle the stability issue [25, 41, 56] (corresponding to the third issue), the above-mentioned methods only utilize the lower order structure of the network while neglecting higher order characteristics. In this work, we aim to address these issues and the upcoming challenges include the following:

(1) How to explore multi-scale connections in the network and leverage them in label propagation?
(2) How to distinguish the strength or the importance of connections beyond the lower order structure in the framework of label propagation?
(3) How to alleviate the randomness in the label propagation process without reducing its efficacy?

3 THE PROPOSED METHOD

In order to address the issues mentioned in the above section, we propose a novel *MWLP* method for community detection. A diagram of the whole framework is shown in Figure 3. Specifically, we propose the following three approaches to tackling these issues:

(1) We investigate the higher order characteristics of the network by identifying the most representative motif, i.e., motif of interest.
(2) A re-weighted network is designed based on the motif of interest, where the strength of connections is encoded.
(3) A novel voting strategy, termed *NaS* (applied to the label propagation) is proposed to alleviate the randomness caused by the tie breaking.

More in detail, for (1), higher order characteristics of the network can be captured by the motif of interest, which is the most representative motif for a network and would be of high statistical significance. What's more, the motif of interest can help to gain new insights into the network organization beyond the lower order structure (at the level of individual nodes and edges) and can reflect a certain kind of phenomenon in the network. For instance, the triangle motif is the motif of interest for most of the social networks, since the *homophily principle* [26] captured by triangles is an intrinsic property of human activities. The ubiquity of triangles is also a reflection of the transitive relations in social networks [11, 31, 32]. And for (2), upon obtaining the motif of interest, we can construct the motif-based hypergraph to encode the higher order characteristics of the

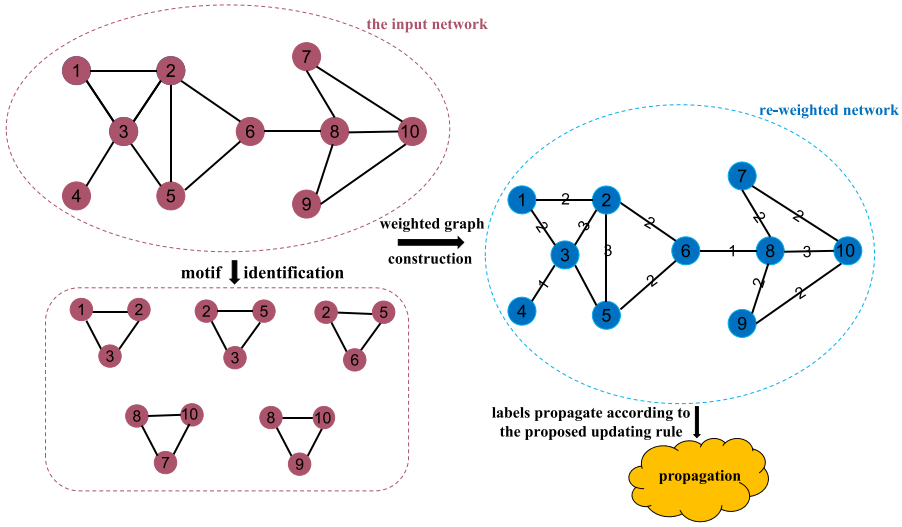


Fig. 3. Illustration of the proposed method, i.e., MWLP (the numbers denote node ids). The re-weighted network is constructed regarding the original connections in the input network and the motif-based higher order connections. Note that the method can work on both weighted and unweighted input network and we just take the unweighted network as an example.

network [22]. By design, a re-weighted[1] network can be constructed by unifying the structure of original graph and the motif-based hypergraph. Under this regime, the strength or the importance of connections can be distinguished. As for (3), we apply the proposed voting strategy, i.e., *NaS* to the label propagation, aiming at reducing the chance that more than one labels gain the highest votes (alleviating the random selection).

In the following, we will detail the motif-based higher order structure mining and the motif-aware weighted label propagation.

3.1 Motif-Based Higher Order Structure Mining

As introduced in Section 2.1, motifs play crucial roles in uncovering structural design principles of networks. Thus, the motif-based higher order structure mining is necessary and important in this sense. Specifically, the mining task consists of the following two subtasks:

(1) Motif identification in the network, that is, finding the motif of interest in the given network.
(2) Uncovering the structural characteristics of the network based on the motifs that have been identified in the given network.

Suppose we are given a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set. Let $n = |\mathcal{V}|$ and $m = |\mathcal{E}|$ denote the number of nodes and edges, respectively. The first subtask aims to find the recurring patterns (or subgraphs) that are statistically overrepresented compared with those in the corresponding random networks. Hence, isomorphic subgraph searching, random network generation, and statistical significance measuring are three important steps when performing the motif identification. For clarity, we define $\mathcal{M}(p, q)$ as the motif with p nodes and q edges. Note

[1]by "re-weighted," we mean that the original network A can be weighted network ($A(i, j) = 1$, $\forall i, j \in \{1, 2, \ldots, n\}$ is simply a trivial case).

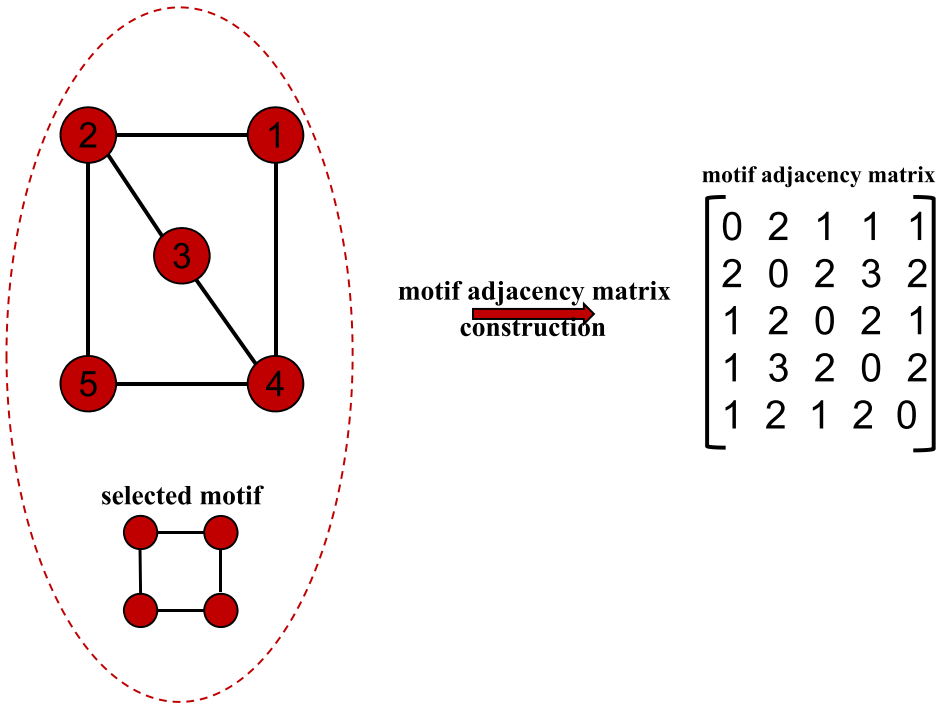


Fig. 4. The construction of the motif adjacency matrix (the numbers denote node ids).

that there can be a set of motif variants with p nodes and q edges, which are not isomorphic to each other, but only the motif of interest will be selected for one specific network as stated above. For example, the identified motif in Figure 4 can be denoted as $\mathcal{M}(4, 4)$ since it consists of 4 nodes and 4 edges. The statistical significance of the motif can be measured by z-score (abbr. z), which is calculated as follows [53]:

$$z(\mathcal{M}(p, q)) = \frac{F_{In} - \overline{F}_{Rnd}}{\sqrt{\sigma^2_{Rnd}}}, \tag{2}$$

where F_{In} stands for the occurrence frequency of the motif in the input network, $\overline{F}_{Rnd}$ and σ^2_{Rnd} stand for the mean and variance of the occurrence frequencies in the corresponding random networks, respectively. Generally, a motif is regarded as "statistically significant" if the associated z-score is larger than 2.0 [53].

Many efforts have been made on motif identification, such as Mfinder [18], MAvisto [36], FANMOD [52], and Kavosh [17]. Even some GPU based approaches have been developed [23]. In our experiments, we utilize *FANMOD* [52] to identify motif of interest according to the z-score and motif with highest z-score would be selected. The *FANMOD* tool implements a novel algorithm called RAND-ESU [50] to enumerate and sample subgraphs. More in detail, the user can specify a certain integer k (usually ranges from 2 to 8) for the motif size, i.e., the number of nodes that involved in the motif, and then different variants of size-k motif can be identified and corresponding frequencies in the original network, mean and standard deviations of frequencies in the random networks as well as the z-scores will be reported. We set $k = 3$ in our experiments since we focus on triangle motifs and the size-3 motif with highest z-score is selected as the motif of interest for the input network.

As for the second subtask, we emphasize the motif-based higher order connections between nodes by introducing the motif adjacency matrix $W_{\mathcal{M}}$ [3], which is the matrix form of the so-called hypergraph in the literature [22]. The motif-based hypergraph encodes higher order characteristics of the network, which helps to gain new insights into the design principles and can be crucial for graph clustering. Formally, the motif adjacency matrix can be calculated as follows:

$$W_{\mathcal{M}}(i,j) = I_{ij}^{\mathcal{M}}, \tag{3}$$

where $I_{ij}^{\mathcal{M}}$ stands for the number of instances of motif $\mathcal{M}$ that contain both node i and node j. Figure 4 shows the construction of the motif adjacency matrix explicitly, where the identified motif is the four-node motif. Note that different networks may have different motifs and may give different interpretations to the identified motifs according to the characteristics of the network (different motifs can model different kinds of relations in the network). In this example, we just take the four-node motif for illustration purpose. In other words, once the motif is identified, we can infer the prevalent types of relations in the network by giving interpretations to the identified motif. Hence, the structural characteristics can be uncovered and the second subtask is completed.

The motif adjacency matrix $W_{\mathcal{M}}$ can provide significantly richer structural information than the conventional node adjacency matrix (based on one-hop connections). For example, as shown in Figure 4, node 1 and node 3 do not have any direct one-hop connections but there are paths between them, which can be captured by the motif structure and encoded in the motif adjacency matrix, i.e., the elements positioned $(1,3)$ and $(3,1)$ are greater than zero, indicating the higher order connections in view of the identified motif. Thus, the first issue mentioned in Section 2.2 can be solved.

What's more, according to the definition, the motif adjacency matrix is a weighted matrix and larger weights imply closer higher order connections based on the identified motif. Intuitively, the intimacy of nodes can be defined through the motif adjacency matrix. However, lower order structural information would be lost if only the higher order structural information is taken into consideration. Therefore, for preserving both higher order and lower order structure, a novel re-weighted network is designed, which can be represented in matrix form as follows:

$$W = A + W_{\mathcal{M}}, \tag{4}$$

where A is the node adjacency matrix for the given network, whose edges can be weighted or unweighted and $W_{\mathcal{M}}$ is the motif adjacency matrix based on the identified motif of interest, i.e., $\mathcal{M}$, which captures the higher order characteristics of the network. In this way, $W(i,j)$ can be regarded as a measure of intimacy (with respect to both lower order and higher order connections) between node i and j, and the strength or the importance of connections can be distinguished by this measure. Thus, the second issue mentioned in Section 2.2 can be solved.

3.2 Motif-Aware Weighted Label Propagation

The label propagation process proceeds over the newly designed re-weighted network. Unlike basic label propagation algorithms, which adopt the majority voting strategy in label updating (i.e., updating the current label of each node to the label shared by most of its neighbors [42]), the proposed MWLP method considers not only the quantity of neighbors but also the quality (strength or importance) of connections with neighbors. To achieve this goal, we propose a voting strategy termed *NaS*, which takes into consideration both the Number and the Strength of connections. Specifically, we assign a voting score to each neighbor who would have influence on the label updating of node v. Denote by $\Gamma(v)$ the set of neighbors (considering both lower order and higher order connections) of node v, and for every $i \in \Gamma(v)$, the voting score can be calculated as

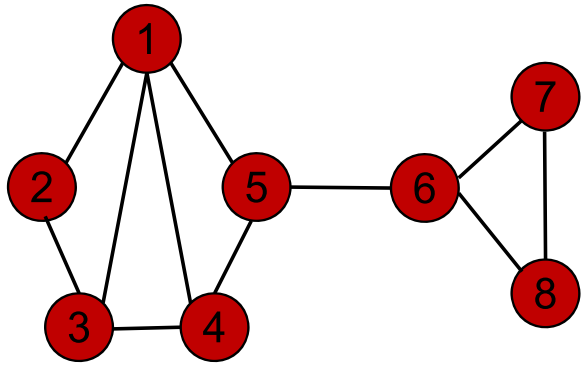


Fig. 5. The initial state of the synthetic network, which is designed to verify the stability of the proposed method (the numbers denote node labels).

follows:

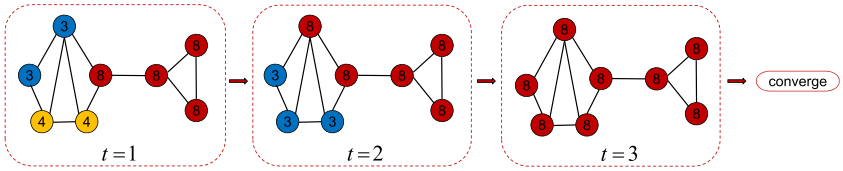
$$S_{vot}(i) = \lambda|\Gamma^{C_i}(v)| + (1-\lambda)W(v,i), \tag{5}$$

where C_i is the label of neighbor i, $|\Gamma^{C_i}(v)|$ denotes the number of neighbors of v with the same label as i, $W(v,i)$ measures the motif-based intimacy between node v and its neighbor i, and λ is a tradeoff parameter that balances the influence between the number of labels of the neighboring nodes and the higher order intimacy of the corresponding nodes. This parameter can be customized for different networks concerning the diverse structures.

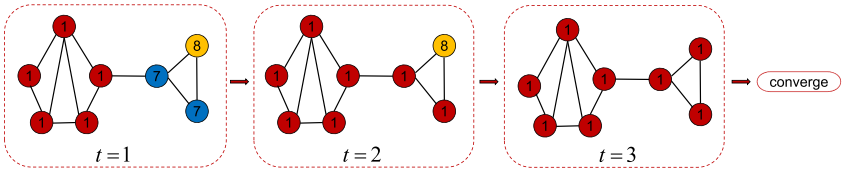
Based on the newly designed voting score, a novel label updating rule can be designed, that is, the label of node v is updated to the label of its neighbor with the highest voting score. If more than one neighbors hold the same highest voting score, they are randomly selected. In fact, the random selection scenario is less likely to happen compared with that in the basic label propagation algorithm since the voting score is determined by two terms, i.e., $|\Gamma^{C_i}(v)|$, $W(v,i)$, as well as the tradeoff parameter λ. Therefore, it becomes more difficult for two or even more nodes to hold the same voting score. Thus, the stability of the method is relatively improved by reducing the randomness and the third issues as mentioned in Section 2.2 can be alleviated. In order to verify the superiority of the proposed method over the RAK method in terms of stability, a synthetic network is designed and each node is initialized with unique label as shown in Figure 5. We run the RAK method and the proposed *MWLP* method several times over the same network setting. The results of 3 runs are recorded and shown in Figures 6 and 7, respectively. As can be seen, the RAK method may give rise to different community structures in different runs over the same network setting while the proposed *MWLP* method is more stable and can discover satisfactory network partitions. The label propagation stops when the label of each node in the network does not change or the maximum number of iterations is reached. And the goal of community detection is achieved by grouping together those nodes with the same label. In conclusion, we perform the motif identification followed by the higher order structure mining to unravel the characteristics of the network, whereby a weighted network is obtained from the original network that captures both lower order and motif-based higher order structural characteristics. In addition, a novel voting strategy is designed for label updating in the label propagation process, by which more stable results can be obtained compared with basic label propagation algorithms.

3.3 Method Summary and Computational Complexity

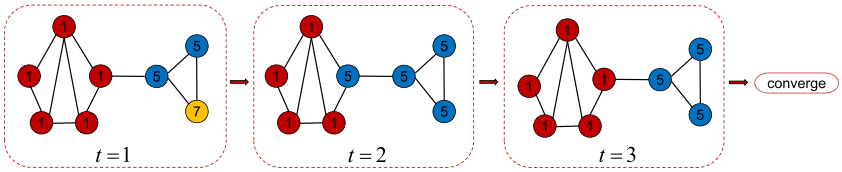
The complexity of the proposed method is governed by the motif-based higher order structure mining, in particular, the construction of the motif adjacency matrix, i.e., $W_{\mathcal{M}}$ (the label



(a) Run 1: Degenerated results are obtained where all nodes are assigned to the same community.



(b) Run 2: Degenerated results are obtained where all nodes are assigned to the same community.



(c) Run 3: Correct results are obtained.

Fig. 6. Results of 3 runs by the RAK method (the numbers denote node labels).

propagation can be accomplished in nearly linear time [33]). Specifically, the computational time is bounded by the time to find all instances of the motif in the network. For a motif with p nodes, we can check each p-tuple of nodes in the graph in $\Theta(n^p)$ time [3], where n is the number of nodes in the network. However, most real-world networks are sparse and there are several efficient algorithms dealing with this case [7, 12, 51, 52]. As for the triangle motif or $\mathcal{M}(3,3)$ as shown in Figure 8(b), which is the prevalent structure especially in social networks [11, 28, 31, 32], there is an algorithm with computational complex $\Theta(m^{1.5})$ to perform the triangle counting [19], where m is the number of edges in the network.

For clarity, Algorithm 1 summarizes the proposed *MWLP* method.

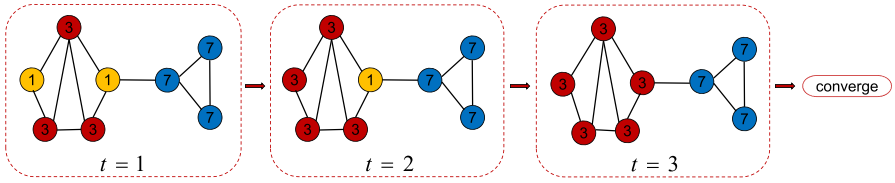
4 EXPERIMENTS

To validate the effectiveness of our method, extensive experiments are conducted on seven real-world datasets. The results demonstrate the superiority of our method over the baseline methods. The implementation of the proposed method is available at https://github.com/lipzh5/MWLP.git

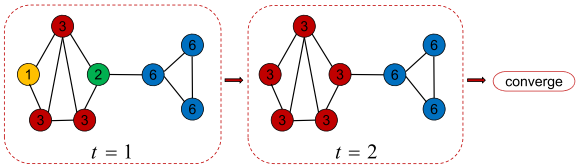
4.1 Dataset Description

Seven datasets obtained from https://linqs.soe.ucsc.edu/data are used, which can be roughly divided into two categories, i.e., webpage citation networks including Cornell, Texas, Washington, and Wisconsin, and scientific publication citation networks including Cora, Citeseer, and Pubmed.

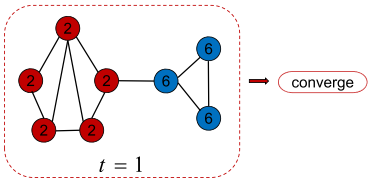
For clarity, the statistics of the datasets are listed in Table 1, where n is the number of nodes, m is the number of edges (all the edges are treated as undirected) and K is the number of ground truth communities in the network. The z-scores of the motif of interest are reported, which are



(a) Run 1: Correct results are obtained.

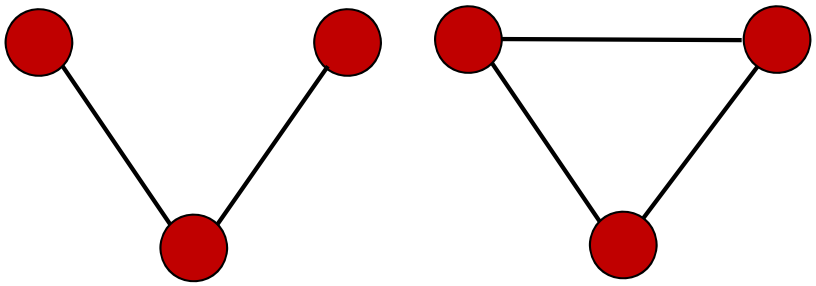


(b) Run 2: Correct results are obtained.



(c) Run 3: Correct results are obtained.

Fig. 7. Results of 3 runs by the proposed MWLP method (the numbers denote node labels). Compared with Figure 6, the proposed MWLP method can improve the performance and generate more stable results.



(a) $\mathcal{M}(3,2)$ (b) $\mathcal{M}(3,3)$

Fig. 8. Illustration of three-node motifs: denote by $\mathcal{M}(3,2)$ and $\mathcal{M}(3,3)$, the motifs with 3 nodes, 2 edges and 3 nodes, 3 edges, respectively.

computed over 1,000 random networks respectively via the *FANMOD* tool [52]. λ_{opt} denotes the optimal tradeoff parameter, i.e., λ that can give rise to better results compared with other parameter settings in the experiments. And $\hat{K}$ is the average number of communities detected over 20 runs by the proposed *MWLP* when λ_{opt} is adopted (a ceiling function will be applied to obtain integers).

Motif-based higher order structure mining is performed on these networks to find representative motifs and then we can give proper interpretations to the identified motifs. Here, we focus on the

Table 1. Summary of Seven Real-World Networks

Datasets	n	m	K	$z(\mathcal{M}(3,3))$	λ_{opt}	$\hat{K}$
Cornell	195	304	5	31.587	0.3	39
Texas	187	328	5	19.765	0.1	31
Washington	230	446	5	26.506	0.1	58
Wisconsin	265	530	5	38.791	0.3	34
Cora	2,708	5,429	7	744.49	0.1	82
Citeseer	3,312	4732	6	153.49	1.0	438
Pubmed	19,717	44,324	3	642.36	0.1	2,522

ALGORITHM 1: Motif-Aware Weighted Label Propagation (MWLP)

Input: Node adjacency matrix $A \in \mathbb{R}^{n\times n}$, the trade-off parameter λ and the maximum number of iterations T.

Initialize the label assignment vector: $\xi^{(t)}(i) = i, \forall i \in \{1, 2, \ldots, n\}$ and set $t = 0$;
Initialize the indicator vector $\delta \in \mathbb{R}^{n\times 1}$ to be a zero vector.
Identify motif of interest, i.e., $\mathcal{M}$.
Construct motif adjacency matrix $W_{\mathcal{M}}$ from A via (3).
Obtain the re-weighted network via (4).
repeat
  Set $t = t + 1$;
  repeat
    Randomly pick one node, e.g. v, that has not been visited, i.e., $\delta(v) = 0$;
    Calculate the voting scores for all the neighbors (including those with higher-order connections) of v according to equation (5);
    Choose the label, e.g., θ of the neighbor with the highest voting score;
    Update $\xi^{(t)}(v) = \theta$ and set $\delta(v) = 1$.
  until $\Sigma_{j=1}^{n}\delta(j) = n$.
until $\xi^{(t)}(v) = \xi^{(t-1)}(v), \forall v \in \{1, 2, \ldots, n\}$ or $t > T$.
Obtain the label assignment vector $\xi = \xi^{(t)}$.

Output: Label assignment vector $\xi \in \mathbb{R}^{n\times 1}$ and the number of detected communities $\hat{K}$.

three-node motifs as shown in Figure 8, but our method extends to other motifs as well. Further analysis will be given on the Cornell dataset as an example.

The closed triangle motif, i.e., $\mathcal{M}(3,3)$ as shown in Figure 8(b) is identified to be of statistical significance for all the seven datasets. Specifically, for Cornell, $z(\mathcal{M}(3,3)) = 31.587$, where $z(\mathcal{M}(3,3))$ stands for the z-score of $\mathcal{M}(3,3)$.

The Cornell dataset is a webpage citation network gathered from the Cornell university, where nodes represent webpages and edges represent the links or webpage accesses. Actually, the underlying objects are corresponding people who may obtain course information or entertainment news by accessing certain webpages. If two people are friends or classmates, they may have similar interests and webpage accesses, which accounts for the closed triangles in the network. Thus, the closed triangle motif, i.e., $\mathcal{M}(3,3)$ can be used to model the friendship relation for the Cornell dataset. Similar analysis can be given on other three webpage citation networks. As for the scientific publication citation networks, the Cora dataset is taken as an example, where 1,630 closed triangles are identified and the corresponding closed triangle motif, i.e., $\mathcal{M}(3,3)$ can model the triangle relations engendered from the scientific cooperation.

4.2 Evaluation Measure

Two commonly used evaluation measures, i.e., normalized mutual information (NMI) [24] and purity [35] are utilized to evaluate the performance of community detection methods:

—*NMI*

$$NMI(C, C^*) = \frac{MI(C, C^*)}{\max(H(C), H(C^*))}, \tag{6}$$

where C and C^* denote the detected communities and the ground truth communities respectively. $H(C)$ is the entropy of the partition C and $MI(C, C^*)$ is the mutual information between C and C^*.

—*Purity*

$$P = \frac{\sum_j \max_c\{\phi_j(c)\}}{n}, \tag{7}$$

where j denotes the index of detected community, c denotes the ground truth or real community label.

$\phi_j(c)$ = number of nodes from real community c occurring in the detected community j and n is the number of nodes in the whole network.

4.3 Baseline Methods

We adopt seven community detection methods as our baseline methods including SLPA [57] fast-Newman [29], Ncut [40], NMF [48] APMOEA [38], TJA-net [37], and the community integration method (abbr. ComInt) [39].

—SLPA: this is a fast algorithm based on the general speaker-listener information propagation process, which spreads a label at a time between nodes according to interaction rules. It does not require any knowledge about the number of community [57].
—Fast-Newman: this is an agglomerative hierarchical clustering method based on modularity, where a "greedy" optimization scheme is taken into consideration to achieve a better performance [29].
—Ncut: this method is most related to the graph theoretic formulation of grouping, where the *normalized cut* is proposed as a measure of the goodness of the partition and the minimization problem is solved to obtain a good partition [40].
—NMF: this method performs the community detection based on NMF, which can be regarded as a powerful tool for data analysis with enhanced interpretability [48].
—APMOEA: this is a multi-objective evolutionary algorithm based on affinity propagation, which combines a powerful data clustering method, i.e., affinity propagation, and an evolutionary algorithm to achieve better detection results within a few iterations [38].
—TJA-net: this method can effectively identify communities in a large network with a good balance between accuracy, stability, and computation time, where the modularity density is proposed and used as the objective function to deal with the issue of resolution limit for different network structures [37].
—ComInt: A community integration strategy is proposed in this method. It initially searches for potential community center and then arranges the pre-partitioned communities. To ensure that communities with greater influence are prioritized during the process of community integration, an improved modularity density increment is proposed [39].

4.4 Comparison Results

The comparison results are shown in Tables 2 and 3 in terms of NMI and purity, respectively. We run the methods 20 times and record the corresponding mean values as well as the standard

Table 2. Results on Seven Real World Datasets, Where the Evaluation Measure is NMI and the Standard Deviations Over 20 Runs Are Also Presented

Methods	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer	Pubmed
MWLP	**0.211**±0.001	0.165±0.000	**0.288**±0.002	0.171±0.002	0.336±0.001	**0.231**±0.001	**0.288**±0.002
SLPA	0.117±0.007	0.067±0.009	0.058±0.010	0.097±0.010	0.278±0.005	0.213±0.001	0.099±0.001
fast-Newman	0.124±0.000	0.070±0.000	0.092±0.000	0.075±0.000	**0.360**±0.000	0.229±0.000	0.156±0.000
Ncut	0.067±0.006	0.028±0.000	0.049±0.000	0.036±0.005	0.118±0.005	0.136±0.003	0.264±0.000
NMF	0.051±0.011	0.077±0.013	0.038±0.008	0.059±0.022	0.331±0.013	0.093±0.008	0.137±0.001
APMOEA	0.025±0.029	0.056±0.058	0.089±0.091	0.058±0.060	0.075±0.122	0.037±0.071	NA
TJA	0.150±0.010	0.107±0.019	0.156±0.011	0.101±0.004	0.280±0.000	0.215±0.000	0.100±0.000
ComInt	0.204±0.001	**0.174**±0.002	0.204±0.001	**0.192**±0.001	0.192±0.000	0.167±0.001	0.059±0.000

The best results on each dataset are highlighted in bold.

Table 3. Results on Seven Real World Datasets, Where the Evaluation Measure is Purity and the Standard Deviations Over 20 Runs Are Also Presented

Methods	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer	Pubmed
MWLP	0.569±0.002	0.626±0.000	0.755±0.002	0.564±0.005	0.769±0.024	0.721±0.002	0.755±0.002
SLPA	0.470±0.007	0.571±0.011	0.517±0.015	0.528±0.010	0.827±0.004	0.779±0.003	0.830±0.001
fast-Newman	0.482±0.000	0.583±0.000	0.570±0.000	0.509±0.000	0.776±0.000	0.755±0.000	0.750±0.000
Ncut	0.433±0.005	0.556±0.000	0.517±0.000	0.479±0.016	0.370±0.005	0.414±0.006	0.680±0.000
NMF	0.429±0.004	0.577±0.014	0.486±0.011	0.487±0.026	0.549±0.014	0.351±0.010	0.580±0.006
APMOEA	0.226±0.228	0.290±0.294	0.339±0.344	0.291±0.296	0.212±0.354	0.127±0.242	NA
TJA	0.483±0.005	0.587±0.017	0.596±0.009	0.503±0.008	**0.855**±0.003	**0.804**±0.000	**0.833**±0.000
ComInt	**0.653**±0.003	**0.739**±0.004	**0.799**±0.002	**0.773**±0.004	0.727±0.001	0.669±0.000	0.709±0.000

The best results on each dataset are highlighted in bold.

Table 4. Average Rank Across Seven Real-World Datasets

Methods	MWLP	SLPA	fast-Newman	Ncut	NMF	APMOEA	TJA	ComInt
Avg.rank (NMI)	1.43	5.14	3.43	6.29	5.57	7.29	3.43	3.43
Avg.rank (purity)	2.71	3.57	3.71	6.29	6.29	7.86	2.43	2.71

deviations (NA means that the result cannot be obtained within 5 days). Additionally, the average ranks are also provided in Table 4, which are computed by averaging the ranking positions of each method across the datasets.

The proposed *MWLP* works well on these real-world datasets especially when NMI is adopted as the evaluation measure. Although it could not achieve the highest score over all the datasets, the average ranks concerning both NMI and purity are always in top three. It even beats all the compared baseline methods with the average NMI rank 1.43 and wins the second place with average purity rank 2.71. In particular, the NMI values of the proposed *MWLP* rank the first on Cornell, Washington, Citeseer, and Pubmed with improvement 3.43%, 41.18%, 0.87%, and 9.09%, respectively, compared with the best baseline method. On the other three datasets, the NMI values of the proposed *MWLP* rank the second. When measured by purity, the *MWLP* method still performs well. Specifically, the purity values on the first four webpage citation networks (i.e., Cornell, Texas, Washington, and Texas) rank the second and on the other three scientific publication citation networks (i.e., Cora, Citeseer, and Pubmed), the purity values are 0.769, 0.721, and 0.755, respectively. Actually, the proposed *MWLP* method suffers a little in purity though

Table 5. The Number of the Detected Communities, i.e., $\hat{K}$ for Methods that do not Take K as Input

Datasets	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer	Pubmed
MWLP	39	31	58	34	82	438	2,522
SLPA	15	12	11	19	233	529	1,266
fast_Newman	2	2	3	2	6	25	113
APMOEA	5	15	28	21	121	116	–
TJA	19	17	28	19	344	704	1457
ComInt	92	91	151	159	1,783	1,825	9,089

it shows superiority in NMI. The key reason may be attributed to the number of the detected communities. For clarity, Table 5 reports the number of the detected communities for methods that do not take the ground truth community number, i.e., K as input. As can be seen, the label propagation-based methods (e.g., *MWLP*, SLPA) are prone to obtain more communities compared with the modularity optimization-based methods (e.g., fast_Newman). Among those methods, ComInt generates the largest number of communities across all the datasets, which may account for the better performance in purity. Notice that more communities may improve the value of purity to some extent. Besides, if we focus more on NMI as shown in Table 2 to alleviate the influence of the detected number of communities in a certain degree, we may find that the proposed *MWLP* method can achieve better performance in larger networks (e.g., Citeseer, Pubmed).

What's more, the improvement on the stability can be reflected from the smaller standard deviations (which indicates more stable performance) in contrast to other label propagation-based methods, i.e., SLPA. The stability is mainly achieved by introducing a novel voting strategy, i.e., *NaS*, and based on the newly designed voting score, the label propagation can proceed in a more stable way. Since the tradeoff parameter λ plays the role of balancing the importance between the number and the strength of the connections and may cast great influence on the community detection, we have tried different settings for this parameter and choose the optimal one for each dataset, which is listed in Table 1. What's more, the proposed method can quickly converge in several iterations and thus is feasible to deal with large-scale networks.

4.5 Parameter Analysis

The effect of the tradeoff parameter, i.e., λ in Equation (5), on the community detection is shown in Figure 9. As can be seen, the performance of community detection (measured by NMI and purity) varies slightly as λ changes, which is especially obvious for the first four webpage citation datasets (see Figure 9(a)–(d) for clarity). As for the other three scientific publication citation networks, the purity values tend to decrease as λ increases, which means better partitioning (measured by purity) can be obtained when the structure of the re-weighted network is endowed with larger weight. The results also confirm the necessity of utilizing the motif-based higher order structure and the effectiveness of the proposed method. However, since λ plays a role of balancing the influence of the number of connections and the strength of the connections concerning the motif structure, the optimal value may be different for different datasets. What's more, for the scientific publication citation networks, the influences of λ are more remarkable (compared with the webpage citation networks) since they are relatively larger and more iterations are needed before the label propagation stops. In other words, the duration of λ acting as a tradeoff factor is longer, and thus setting different values for λ would be more likely to produce various results. As for how to select an appropriate λ for a given network where the ground truth is not accessible, we can make use of the structural measures such as modularity as well as domain knowledge to choose an optimal value with higher modularity and a reasonable number of communities.

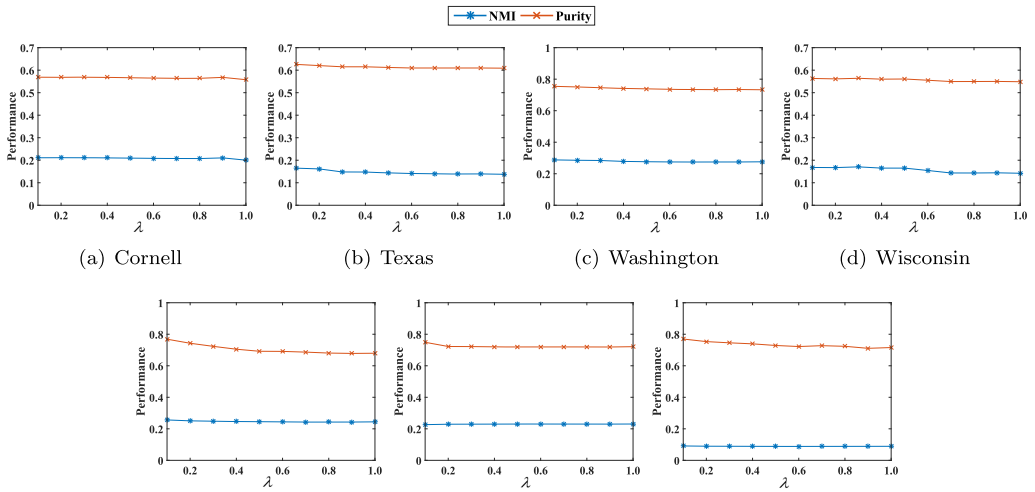


Fig. 9. Parameter analysis of λ: the evaluation measures are NMI and purity.

5 CONCLUSION

In this article, we propose a *MWLP* method for community detection. The proposed method can shed light on the higher order structure of the network by motif mining. In addition, by seamlessly integrating the motif-based higher order features with the lower order structure of the network, we distinguish different connections (i.e., consider not only the number but also the strength of connections) between nodes and a novel voting strategy termed *NaS* is proposed for label updating. Although we focus on triangle motifs in this work, many other interesting patterns can also be explored and utilized in the future. What's more, the proposed method is more stable than the basic label propagation algorithms owing to the design of the re-weighted network as well as the novel voting strategy. Extensive experiments on seven real-world datasets demonstrate that the proposed method achieves better performance over the baseline methods.

REFERENCES

[1] Alex Arenas, Alberto Fernandez, Santo Fortunato, and Sergio Gomez. 2008. Motif-based communities in complex networks. *Journal of Physics A: Mathematical and Theoretical* 41, 22 (2008), 224001.

[2] Punam Bedi and Chhavi Sharma. 2016. Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 6, 3 (2016), 115–135.

[3] Austin R. Benson, David F. Gleich, and Jure Leskovec. 2016. Higher-order organization of complex networks. *Science* 353, 6295 (2016), 163–166.

[4] Oualid Boutemine and Mohamed Bouguessa. 2017. Mining community structures in multidimensional networks. *ACM Transactions on Knowledge Discovery from Data* 11, 4 (2017), 51.

[5] Tanmoy Chakraborty, Sriram Srinivasan, Niloy Ganguly, Animesh Mukherjee, and Sanjukta Bhowmick. 2016. Permanence and community structure in complex networks. *ACM Transactions on Knowledge Discovery from Data* 11, 2 (2016), 14.

[6] Zheng Chen, Xinli Yu, Bo Song, Jianliang Gao, Xiaohua Hu, and Wei-Shih Yang. 2017. Community-based network alignment for large attributed network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 587–596.

[7] Sofie Demeyer, Tom Michoel, Jan Fostier, Pieter Audenaert, Mario Pickavet, and Piet Demeester. 2013. The index-based subgraph matching algorithm (ISMA): Fast subgraph enumeration in large networks using optimized search trees. *PLOS One* 8, 4 (2013), e61183.

[8] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. 2002. Self-organization and identification of web communities. *Computer* 35, 3 (2002), 66–70.

[9] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3–5 (2010), 75–174.
[10] Michelle Girvan and Mark E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99, 12 (2002), 7821–7826.
[11] Paul W. Holland and Samuel Leinhardt. 1977. A method for detecting structure in sociometric data. In *Social Networks*. Elsevier, 411–432.
[12] Maarten Houbraken, Sofie Demeyer, Tom Michoel, Pieter Audenaert, Didier Colle, and Mario Pickavet. 2014. The index-based subgraph matching algorithm with general symmetries (ISMAGS): Exploiting symmetry for faster subgraph enumeration. *PLOS One* 9, 5 (2014), e97896.
[13] Ling Huang, Hong-Yang Chao, and Guangqiang Xie. 2020. MuMod: A micro-unit connection approach for hybrid-order community detection. In *Proceedings of the AAAI*.
[14] Ling Huang, Chang-Dong Wang, and Hong-Yang Chao. 2018. A harmonic motif modularity approach for multi-layer network community detection. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'18)*. 1043–1048.
[15] Ling Huang, Chang-Dong Wang, and Hong-Yang Chao. 2019. Higher-order multi-layer community detection. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19), the 31st Innovative Applications of Artificial Intelligence Conference (IAAI'19), the 9th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'19)*. 9945–9946.
[16] Ling Huang, Chang-Dong Wang, and Hong-Yang Chao. In Press 2019. oComm: Overlapping community detection in multi-view brain network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
[17] Zahra Razaghi Moghadam Kashani, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz Saberi Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. 2009. Kavosh: A new algorithm for finding network motifs. *BMC Bioinformatics* 10, 1 (2009), 318.
[18] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. 2004. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* 20, 11 (2004), 1746–1758.
[19] Matthieu Latapy. 2008. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science* 407, 1–3 (2008), 458–473.
[20] Pei-Zhen Li, Yue-Xin Cai, Chang-Dong Wang, Mao-Jin Liang, and Yi-Qing Zheng. 2019. Higher-order brain network analysis for auditory disease. *Neural Processing Letters* 49, 3 (2019), 879–897.
[21] Pei-Zhen Li, Ling Huang, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. 2018. Community detection using attribute homogenous motif. *IEEE Access* 6 (2018), 47707–47716.
[22] Pei-Zhen Li, Ling Huang, Chang-Dong Wang, and Jian-Huang Lai. 2019. EdMot: An edge enhancement approach for motif-aware community detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*. 479–487.
[23] Wenqing Lin, Xiaokui Xiao, Xing Xie, and Xiaoli Li. 2017. Network motif discovery: A GPU approach. *IEEE Transactions on Knowledge and Data Engineering* 29, 3 (2017), 513–528.
[24] Liyuan Liu, Linli Xu, Zhen Wangy, and Enhong Chen. 2015. Community detection based on structure and content: A content propagation perspective. In *Proceedings of the 2015 IEEE International Conference on Data Mining*. IEEE, 271–280.
[25] Hao Lou, Shenghong Li, and Yuxin Zhao. 2013. Detecting community structure using label propagation with weighted coherent neighborhood propinquity. *Physica A: Statistical Mechanics and its Applications* 392, 14 (2013), 3095–3105.
[26] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27, 1 (2001), 415–444.
[27] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: Simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.
[28] Mark E. J. Newman. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98, 2 (2001), 404–409.
[29] Mark E. J. Newman. 2004. Fast algorithm for detecting community structure in networks. *Physical Review E* 69, 6 (2004), 066133.
[30] Mark E. J. Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69, 2 (2004), 026113.
[31] Mark E. J. Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E* 68, 3 (2003), 036122.
[32] Arnau Prat-Pérez, David Dominguez-Sal, Josep-M Brunat, and Josep-Lluis Larriba-Pey. 2016. Put three and three together: Triangle-driven community detection. *ACM Transactions on Knowledge Discovery from Data* 10, 3 (2016), 22.
[33] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76, 3 (2007), 036106.

[34] Maryam Ramezani, Ali Khodadadi, and Hamid R. Rabiee. 2018. Community detection using diffusion information. *ACM Transactions on Knowledge Discovery* 12, 2, Article 20 (Jan. 2018), 22 pages. DOI: https://doi.org/10.1145/3110215

[35] Nachiketa Sahoo, Jamie Callan, Ramayya Krishnan, George Duncan, and Rema Padman. 2006. Incremental hierarchical clustering of text documents. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. ACM, 357–366.

[36] Falk Schreiber and Henning Schwöbbermeyer. 2005. MAVisto: A tool for the exploration of network motifs. *Bioinformatics* 21, 17 (2005), 3572–3574.

[37] Ronghua Shang, Huan Liu, Licheng Jiao, and Amir M. Ghalamzan Esfahani. 2017. Community mining using three closely joint techniques based on community mutual membership and refinement strategy. *Applied Soft Computing* 61 (2017), 1060–1073.

[38] Ronghua Shang, Shuang Luo, Weitong Zhang, Rustam Stolkin, and Licheng Jiao. 2016. A multiobjective evolutionary algorithm to find community structures based on affinity propagation. *Physica A: Statistical Mechanics and its Applications* 453 (2016), 203–227.

[39] Ronghua Shang, Weitong Zhang, Licheng Jiao, Rustam Stolkin, and Yu Xue. 2017. A community integration strategy based on an improved modularity density increment for large-scale networks. *Physica A: Statistical Mechanics and Its Applications* 469 (2017), 471–485.

[40] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (2000), 888–905.

[41] Chang Su, Xiaotao Jia, Xianzhong Xie, and Yue Yu. 2015. A new random-walk based label propagation community detection algorithm. In *Proceedings of the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 1. IEEE, 137–140.

[42] Lovro Šubelj and Marko Bajec. 2011. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Physical Review E* 83, 3 (2011), 036103.

[43] Bing-Jie Sun, Huawei Shen, Jinhua Gao, Wentao Ouyang, and Xueqi Cheng. 2017. A non-negative symmetric encoder-decoder approach for community detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 597–606.

[44] Chao Tong, Jianwei Niu, Jinming Wen, Zhongyu Xie, and Fu Peng. 2015. Weighted label propagation algorithm for overlapping community detection. In *Proceedings of the 2015 IEEE International Conference on Communications (ICC'15)*. IEEE, 1238–1243.

[45] Charalampos E. Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. 2017. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1451–1460.

[46] Chang-Dong Wang, Jian-Huang Lai, and Philip S. Yu. 2013. Dynamic community detection in weighted graph streams. In *Proceedings of the 13th SIAM International Conference on Data Mining*. 151–161.

[47] Chang-Dong Wang, Jian-Huang Lai, and Philip S. Yu. 2014. NEIWalk: Community discovery in dynamic content-based networks. *IEEE Transactions on Knowledge and Data Engineering* 26, 7 (2014), 1734–1748.

[48] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. 2011. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery* 22, 3 (2011), 493–521.

[49] Yue Wang, Xun Jian, Zhenhua Yang, and Jia Li. 2017. Query optimal k-plex based community in graphs. *Data Science and Engineering* 2, 4 (2017), 257–273.

[50] Sebastian Wernicke. 2005. A faster algorithm for detecting network motifs. In *Proceedings of the International Workshop on Algorithms in Bioinformatics*. Springer, 165–177.

[51] Sebastian Wernicke. 2006. Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 3, 4 (2006), 347–359..

[52] Sebastian Wernicke and Florian Rasche. 2006. FANMOD: A tool for fast network motif detection. *Bioinformatics* 22, 9 (2006), 1152–1153.

[53] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. 2011. Biological network motif detection: Principles and practice. *Briefings in Bioinformatics* 13, 2 (2011), 202–215.

[54] Jierui Xie, Mingming Chen, and Boleslaw K. Szymanski. 2013. LabelrankT: Incremental community detection in dynamic networks via label propagation. In *Proceedings of the Workshop on Dynamic Networks Management and Mining*. ACM, 25–32.

[55] Jierui Xie and Boleslaw K. Szymanski. 2011. Community detection using a neighborhood strength driven label propagation algorithm. In *Proceedings of the 2011 IEEE Network Science Workshop*. IEEE, 188–195.

[56] J. Xie and B. K. Szymanski. 2013. LabelRank: A stabilized label propagation algorithm for community detection in networks. In *Proceedings of the 2013 IEEE 2nd Network Science Workshop (NSW'13)*. 138–143. DOI: https://doi.org/10.1109/NSW.2013.6609210

[57] Jierui Xie, Boleslaw K. Szymanski, and Xiaoming Liu. 2011. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 344–349.

[58] Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. 2014. Revealing the hidden language of complex networks. *Scientific Reports* 4 (2014), 4547.

[59] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. 2017. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 555–564.

[60] Han Zhang, Chang-Dong Wang, Jian-Huang Lai, and Philip S. Yu. 2019. Community detection using multilayer edge mixture model. *Knowledge and Information Systems* 60, 2 (2019), 757–779.

Received September 2018; revised October 2019; accepted November 2019