EdMot: An Edge Enhancement Approach for Motif-aware Community Detection

Pei-Zhen Li, Ling Huang, Chang-Dong Wang, Jian-Huang Lai
School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
Guangdong Province Key Laboratory of Computational Science, Guangzhou, China
Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
sysuLiPeizhen@163.com,huanglinghl@hotmail.com,changdongwang@hotmail.com,stsljh@mail.sysu.edu.cn

ABSTRACT

Network community detection is a hot research topic in network analysis. Although many methods have been proposed for community detection, most of them only take into consideration the lower-order structure of the network at the level of individual nodes and edges. Thus, they fail to capture the higher-order characteristics at the level of small dense subgraph patterns, e.g., motifs. Recently, some higher-order methods have been developed but they typically focus on the motif-based hypergraph which is assumed to be a connected graph. However, such assumption cannot be ensured in some real-world networks. In particular, the hypergraph may become fragmented. That is, it may consist of a large number of connected components and isolated nodes, despite the fact that the original network is a connected graph. Therefore, the existing higher-order methods would suffer seriously from the above fragmentation issue, since in these approaches, nodes without connection in hypergraph can't be grouped together even if they belong to the same community. To address the above fragmentation issue, we propose an Edge enhancement approach for Motif-aware community detection (**EdMot**). The main idea is as follows. Firstly, a motif-based hypergraph is constructed and the top K largest connected components in the hypergraph are partitioned into modules. Afterwards, the connectivity structure within each module is strengthened by constructing an edge set to derive a clique from each module. Based on the new edge set, the original connectivity structure of the input network is enhanced to generate a rewired network, whereby the motif-based higher-order structure is leveraged and the hypergraph fragmentation issue is well addressed. Finally, the rewired network is partitioned to obtain the higher-order community structure. Extensive experiments have been conducted on eight real-world datasets and the results show the effectiveness of the proposed method in improving the community detection performance of state-of-the-art methods.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330882

CCS CONCEPTS

• **Information systems** → **Clustering**; *Network data models*; • **Networks** → **Topology analysis and generation**; *Network architectures*;

KEYWORDS

Community detection; Higher-order; Motif; Edge enhancement; Fragmentation issue

ACM Reference Format:
Pei-Zhen Li, Ling Huang, Chang-Dong Wang, Jian-Huang Lai. 2019. EdMot: An Edge Enhancement Approach for Motif-aware Community Detection. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330882

1 INTRODUCTION

Recent years have witnessed a growing trend in many different disciplines that model and interpret structured data as networks [1, 19]. As ubiquitous abstractions depicting relationships among entities, networks have become a focus of data science and network analysis has attracted an increasing amount of attention from different fields such as physics, biology, mathematics and computer science [14, 23, 24]. Community detection is an important task in network analysis that aims to partition the network into communities of strongly connected nodes.

Although many community detection methods have been proposed [5, 10, 31], most of them only take into consideration the lower-order structure of the network at the level of individual nodes and edges, which fails to unravel the higher-order organization of the network. Recently, to go beyond the lower-order connectivity patterns, some motif-based higher-order community detection methods have been proposed [3, 13, 15, 20, 21, 34], which can capture the motif-based characteristics and gain new insights into the organization of the network. However, they typically focus on only the higher-order connections that are encoded in the motif-based hypergraph but underestimate and even violate the original lower-order topological structure. In particular, the hypergraph is assumed to be a connected graph in which the consequent partitioning procedure can be applied. However,

in some real-world networks, such assumption cannot be ensured. That is, the hypergraph may be fragmented to a large number of connected components (with various sizes) and isolated nodes. This is because the higher-order connections among nodes are built upon motifs. Two nodes are said to have higher-order connection if they have involved in at least one common motif and vise versa. It's worth noting that two nodes can be separated in the hypergraph of higher-order connections despite that they are connected in the original network. In this way, the number of connected components would increase and more isolated nodes would appear. These isolated nodes (from the perspective of higher-order connections) will render the community structure with instability. Therefore, the existing higher-order methods would suffer seriously from the above fragmentation issue since nodes without higher-order connections can't be grouped together even if they belong to the same community.

To address the above fragmentation issue, we propose an Edge enhancement approach for Motif-aware community detection (**EdMot**). The main idea is as follows. Firstly, a motif-based hypergraph is constructed and the top K largest connected components (measured by the number of contained nodes) in the hypergraph are partitioned into modules. Afterwards, the connectivity structure within each module is strengthened by constructing an edge set to derive a clique from each module. Based on the new edge set, the original connectivity structure of the input network is enhanced to generate a rewired network, whereby both the higher-order structure and lower-order structure are integrated. Finally, by partitioning the rewired network, the higher-order community structure can be discovered. In this study, we focus on the triangle motif for its ubiquitousness in social networks [12, 26, 29], but our technique can be extended to other motifs as well. Extensive experiments have been conducted on eight real-world datasets and the results show the effectiveness of the proposed method in improving the community detection performance of state-of-the-art methods.

We summarize the main contributions as follows:

1) We present and formalize the hypergraph fragmentation issue suffered by the existing motif-based community detection methods, where higher-order connections are preserved but the original lower-order structure may be underestimated and even violated.
2) We propose an Edge enhancement approach for Motif-aware community detection (**EdMot**), which can not only leverage higher-order connections of the network but also overcome the hypergraph fragmentation issue.
3) Extensive experiments are conducted on eight real-world datasets to show the effectiveness of the proposed method.

2 BACKGROUND AND PROBLEM STATEMENT

Assume that we are given a network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_i | i = 1, \ldots, n\}$ represents the node set consisting of n nodes and $\mathcal{E} = \{e_i | i = 1, \ldots, m\}$ represents the edge set consisting of m undirected and unweighted edges. A node adjacency matrix $A \in \mathbb{R}^{n \times n}$ is used to encode the node-wise connection of the network, which is also known as the lower-order structure of the network [3]. In this paper, the largest connected component of the original network will be extracted and utilized if it contains isolated nodes. We aim to infer the corresponding higher-order structure, which may characterize the building block of the network. In particular, the typical higher-order connectivity patterns, i.e., motifs, are identified to gain new insights into the organization of the network. Formally, a network motif with p nodes and q edges can be denoted as:

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

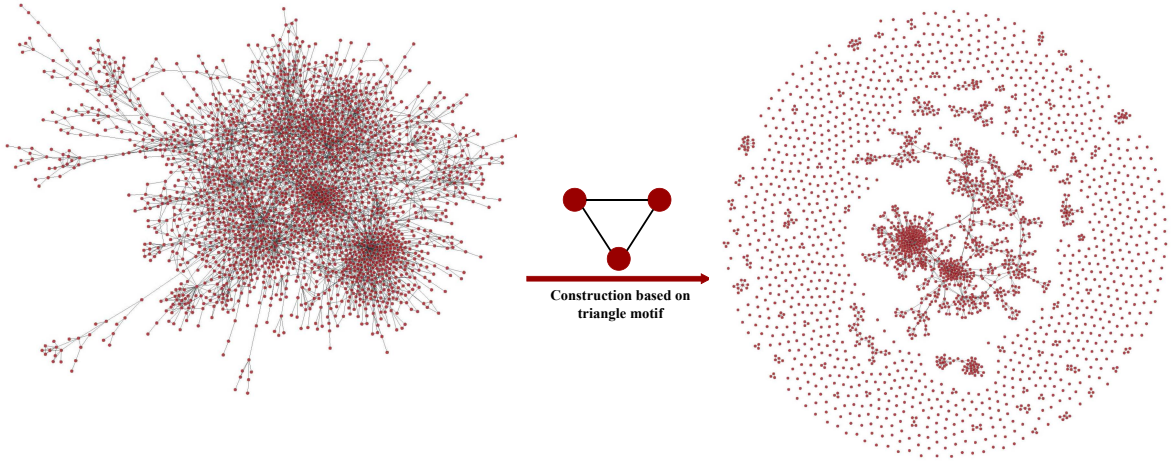
where $\mathcal{V}_\mathbf{M} \subseteq \mathcal{V}$ represents the set of p nodes and $\mathcal{E}_\mathbf{M} \subseteq \mathcal{E}$ represents the set of q edges. Following the convention of the literature [12, 26, 29], we focus on $\mathbf{M}_3^3$, i.e., the triangle motif (as shown in the middle part of Figure 1). However, our technique can be well extended to other motifs. Conceptually, given the original node adjacency matrix A, the motif adjacency matrix constructed from the triangle motif, denoted as $W_\mathbf{M} \in \mathbb{R}^{n \times n}$, can be defined:

$$(W_\mathbf{M})_{ij} = \text{number of motif instances containing nodes } i \text{ and } j. \tag{2}$$

Note that $W_\mathbf{M}$ encodes the motif-based higher-order connections of the network, where edges correspond to co-occurrences in motifs and $(W_\mathbf{M})_{ij}$ can be 0 even though $A_{ij} > 0$.

Several methods have been developed in terms of motif-based community detection [3, 34]. The common denominator of these methods is the dedication to leverage higher-order network structures effectively for community detection. However, they rely on the construction of a hypergraph whose edges correspond to motifs [3]. In this way, the motif-based higher-order structure of the network is highlighted and well preserved but the original lower-order structure is underestimated and even violated. That is, edges that do not involve in any motifs would be eliminated. This leads to the serious issue called hypergraph fragmentation.

By "fragmentation", it means that the hypergraph or the motif adjacency matrix constructed from the original single connected component is usually fragmented into a large number of connected components with various sizes and isolated nodes due to the lack of higher-order connection, i.e., lack of involvement in the motif [37]. As shown in Figure 1, by constructing the motif-based hypergraph from the original network structure of the Cora dataset, a large number of connected components with various sizes and isolated nodes having no connection to any other nodes will be generated from the original single large connected component. As a consequence, most of the existing higher-order community detection methods that utilize the hypergraph would suffer from the hypergraph fragmentation issue [3, 34, 37]. In particular, by definition of communities (nodes should be densely connected within communities but sparsely connected between communities), the existing methods assume that the intermediate hypergraph is a single connected component [3, 34], from which the eventual higher-order community



(a) The original network structure of the Cora dataset (b) Motif-based hypergraph of the Cora dataset

Figure 1: Illustration of the hypergraph fragmentation issue on the Cora dataset: The motif-based hypergraph constructed from the original network consists of several connected components (with various sizes) and a large number of isolated nodes.

structure can be discovered. However, as discussed above, the hypergraph is usually fragmented with a large number of connected components and isolated nodes. Therefore, the performance of the existing higher-order community detection methods can not be guaranteed. For example, there will be a large number of dead nodes in the random walk process. Such isolated nodes would present confusing or rather, generate uncertain community memberships if the random label assignment strategy or the exclusion strategy is applied [16], leading to the degenerate performance.

A naive solution is directly grouping the connected components or isolated nodes together in the hypergraph to form the communities. Unfortunately, these components may be of different sizes and nodes with different ground-truth community labels may reside in the same component. For example, in Figure 1(b), the largest connected component in the hypergraph consists of nodes from 7 different ground-truth communities. Furthermore, serious randomness would be introduced by the isolated nodes due to the lack of higher-order connections because there is no such community definition customized for isolated nodes. In this case, community memberships will be rendered with randomness if the isolated nodes are grouped randomly.

To address the above hypergraph fragmentation issue, we propose a novel method termed **EdMot** for motif-aware community detection. On the one hand, the motif-based higher-order structure is leveraged. On the other hand, the hypergraph fragmentation issue suffered by the traditional higher-order community detection methods can be well addressed by the newly designed edge enhancement strategy.

3 THE PROPOSED METHOD

3.1 Connected Component Identification

We commence by constructing a motif adjacency matrix, (a.k.a. hypergraph) $W_{\mathbf{M}}$ through the original network structure to encode the higher-order connections. It can also be represented in the set form as follows:

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

In the above equation, $\mathcal{G}^{\mathbf{M}}$ represents the motif-based hypergraph, $\mathcal{V}$ represents the node set that is the same as the original network, and $\mathcal{E}^{\mathbf{M}}$ represents the edge set consisting of m_w weighted edges generated based on triangle motifs:

$$\mathcal{E}^{\mathbf{M}} = \{(a, b, \tau)_i\} \tag{4}$$

where $a, b \in \mathcal{V}$ are two end nodes of the i-th edge ($i \in \{1, \cdots, m_w\}$) and τ is the edge weight, i.e., the number of motif instances that contain node a and node b together. Accordingly, a set of c_Φ connected components of the hypergraph can be identified:

$$\Phi = \{\phi_i\} \tag{5}$$

where ϕ_i is the i-th connected component. That is, $\phi_i = \{\mathcal{V}_i^\phi, \mathcal{E}_i^{\mathbf{M}}\}$, $i \in \{1, \cdots, c_\Phi\}$, where $\mathcal{V}_i^\phi \subseteq \mathcal{V}$ is the set of nodes involved in the i-th connected component and $\mathcal{E}_i^{\mathbf{M}} \subseteq \mathcal{E}^{\mathbf{M}}$ is the weighted edge set of the i-th connected component respectively.

Apart from c_Φ connected components, there is a set of isolated nodes in the hypergraph, denoted as

$$\mathcal{V}_{iso} = \mathcal{V} - \bigcup_{i \in \{1, \cdots, c_\Phi\}} \mathcal{V}_i^\phi \tag{6}$$

Actually, motif structures are well preserved in these components, i.e., triangle motifs can certainly be found in the components containing 3 or more nodes. The top K largest connected components (measured by the number of contained nodes), i.e., $\Phi_K \subseteq \Phi$ ($K < c_\Phi$) will be obtained. And each connected component in Φ_K will be further partitioned into modules[1] by using some traditional graph partitioning methods. The influence of the parameter K will be analyzed in later section.

3.2 Connected Component Partitioning

As an example, we adopt the Louvain [5] method that heuristically maximizes the well-known community structure evaluation measure called modularity [24, 25] to partition each connected component $\phi_l \in \Phi_K$ into modules. In particular, by taking each connected component ϕ_l as the input network, the modularity Q can be given as [24]:

$$Q = \frac{1}{4\mu}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2\mu})(s_i s_j + 1) = \frac{1}{4\mu}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2\mu})s_i s_j \tag{7}$$

where k_i and k_j are the degrees of nodes i and j respectively and $\mu = \frac{1}{2}\sum_i k_i$ is the total number of edges in the network. s_i is the community label of node i. $\frac{k_i k_j}{2\mu}$ is the expected number of edges between node i and node j in the randomly rewired graph preserving the same degree distribution. $s_i s_j$ is equal to 1 if node i and node j belong to the same community and -1 otherwise.

The output of the above modularity-based community detection procedure is the partitions (modules) in the l-th connected component $\phi_l \in \Phi_K$. By putting all the partitions (modules) of all the top K largest connected components together, we can obtain a module set, denoted as $\{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}$, where $\bar{m}$ is the number of modules obtained by partitioning all the top K largest connected components. It is worthy noting that many other graph partitioning schemes can also be applied.

3.3 Network Rewiring via Edge Enhancement

The module set $\{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}$ obtained in the previous subsection partially encodes the higher-order community structure. The reason for using "partially" is as follows.

1) **Property I**: If two nodes belong to the same module, it is likely that they have the same ground-truth community label as shown by the existing higher-order community detection approaches [2, 3], i.e., discovering higher-order communities from each connected component in the hypergraph. However, it is also possible that two different modules may belong to the same ground-truth community.
2) **Property II**: The module set $\{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}$ obtained so far involves only part of nodes in the network since only the top K largest connected components in the hypergraph are processed in the connected component partitioning.

To completely reveal the higher-order community structure, the lower-order structure should be taken into account as a complement to the higher-order connections. To this end, an edge enhancement approach is developed to rewire the connectivity structure of the original network, whereby both the connectivity structure within each module in $\{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}$ is strengthened and the original lower-order structure is considered. In this way, not only the higher-order connections of the network can be leveraged but also the hypergraph fragmentation issue can be well addressed.

First of all, the connectivity structure within each module $\{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}$ is strengthened as follows. For the nodes that have already been partitioned into the same module, e.g., $\mathcal{M}_i \in \{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}(i \in \{1, \cdots, \bar{m}\})$, their connectivity is strengthened in line with the assumption that motif-based connection should have higher priority compared with the lower-order connection. This is because motif-based connection reflects the social transitivity and may encode more impressive characteristics of the network [33, 35]. To this end, a clique is obtained for each module $\mathcal{M}_i \in \{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}$ by constructing an edge to each node pair in $\mathcal{M}_i$. In this way, a new set of edges is constructed, denoted as $\mathcal{E}^*_{mod}$,

$$\mathcal{E}^*_{mod} = \{(a, b) | \forall a, b \in \mathcal{M}_i, \forall i = 1, \cdots, \bar{m}\} \tag{8}$$

Notice that, the nodes within each module $\mathcal{M}_i$ are interconnected with each other by the strongest connectivity pattern, i.e. a clique structure, which is almost impossible to be destroyed in the consequent partitioning procedure. According to **Property I**, it is rational to establish such strong connectivity. However, according to **Property II**, it is also necessary to take into account the nodes residing out of the module set as well as the original lower-order connectivity pattern to overcome the fragmentation issue.

Therefore, based on the new edge set $\mathcal{E}^*_{mod}$, the original connectivity structure of the input network is enhanced to generate a rewired network

$$\mathcal{G}^{\mathbf{M}}_A = \{\mathcal{V}, \mathcal{E}^{\mathbf{M}}_A\} \text{ with } \mathcal{E}^{\mathbf{M}}_A = \mathcal{E} \cup \mathcal{E}^*_{mod} \tag{9}$$

In this way, the edge enhanced network contains the same node set as the original network but encodes the connectivity patterns from both the original lower-order network structure in terms of $\mathcal{E}$ and the higher-order connections in terms of $\mathcal{E}^*_{mod}$.

3.4 Method Summary and Computational Complexity

The rewired network, i.e., $\mathcal{G}^{\mathbf{M}}_A$, is fed into some graph partitioning methods to obtain the final community structure, i.e., $\widehat{\mathcal{C}} = \{\mathcal{C}_1, \cdots, \mathcal{C}_c\}$, where c is the number of communities in the final partitioning.

For clarification, Algorithm 1 summarizes the main procedure of the proposed **EdMot** method. In addition, Figure 2 exhibits the whole process intuitively.

We now analyze the computational complexity of the proposed **EdMot** method. Overall, the complexity of the method

[1] To avoid confusing the partitioning results in each connected component with the eventual partitioning results, the "module" is used to name the partitioning results in each connected component.

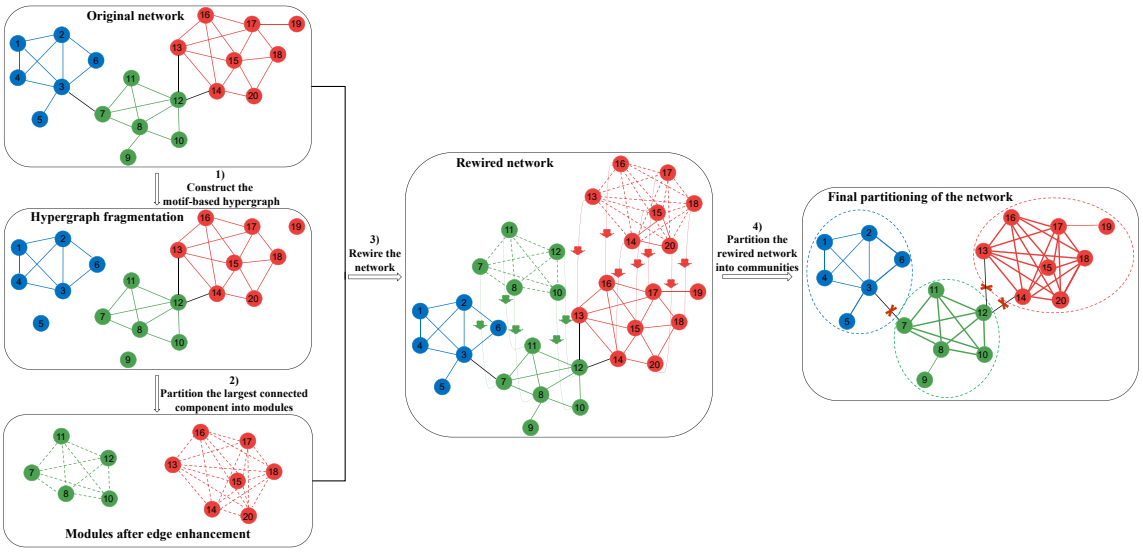


Figure 2: Illustration of the proposed EdMot algorithm. A synthetic network is designed to serve as the original network, where nodes and edges in three communities are denoted with different colors and the black edges represent the inter-community edges. Specifically, by constructing the motif-based hypergraph in step 1, the hypergraph fragmentation issue arises, where two connected components and three isolated nodes are generated in the hypergraph. By partitioning the largest connected component into modules in step 2, two modules can be identified and a new edge set is constructed to derive a clique from each module, as shown as the dashed line. By rewiring the network in step 3, a rewired network can be obtained by substituting the new edge set into the original network. Finally, by partitioning the rewired network into communities, the community structure can be discovered.

Algorithm 1 The proposed **EdMot** approach.

Input: Original network $A \in \mathbb{R}^{n\times n}$, parameter K, a graph partitioning method $\mathtt{S}$.

1: Construct motif adjacency matrix $W_{\mathbf{M}}$ from A via (2).
2: Obtain connected components Φ via (5) and select top K largest connected components $\Phi_K \subseteq \Phi$.
3: Apply $\mathtt{S}$ to partition each $\phi_l \in \Phi_K$, $\forall l \in \{1, \cdots, K\}$ and aggregate the results to form the module set $\{\mathcal{M}_1, \cdots, \mathcal{M}_{\bar{m}}\}$.
4: Construct a new edge set $\mathcal{E}^*_{mod}$ via (8).
5: Rewire the original network to obtain $\mathcal{G}_A^{\mathbf{M}}$ via (9).
6: Feed $\mathcal{G}_A^{\mathbf{M}}$ into $\mathtt{S}$ to obtain final community structure $\widehat{\mathcal{C}}$.

Output: Community structure $\widehat{\mathcal{C}}$.

is governed by the computations of the motif adjacency matrix $W_{\mathbf{M}}$ and the graph partitioning in both the connected components and rewired network. For simplicity, we assume that we can access the edges in a graph in $O(1)$ time and can access and modify the elements in the matrix in $O(1)$ time. The computational time of constructing $W_{\mathbf{M}}$ is bounded by the time to find all the motif instances in the graph. Theoretically, we can compute $W_{\mathbf{M}}$ in $O(n^p)$ time for a motif with p nodes. However, most real-world networks are sparse, we can instead focus on the computational complexity in terms of the number of edges in the network. In particular, for the triangle motifs discussed in this paper, the motif instances can be found in $O(m^{1.5})$ time [3, 4, 18]. As for the graph partitioning, in this paper, we adopt the heuristic modularity maximization method as an example, which can be finished in $O(n \log n)$ on average for its ability to find the hierarchical community structure [5]. Therefore, the overall computational complexity is $O(m^{1.5} + n \log n)$.

4 EXPERIMENTS

In this section, extensive experiments are conducted to confirm the effectiveness of the proposed method. The Matlab code is available in https://github.com/lipzh5/EdMot_pro.git.

4.1 Datasets and Evaluation Measures

In our experiments, eight real-world datasets are used (the edges are treated to be undirected for all the datasets). On the first four datasets, the ground-truth community labels are provided for evaluation purpose and on the rest four datasets, there is no ground-truth community information where the internal evaluation measures such as modularity are used for evaluation.

Datasets with ground-truth community labels:

1) **polbooks**[1]**.** A network of books about US politics, where edges between books represent frequent copurchasing of books by the same buyers. It consists of 105 nodes and 441 edges.

2) **email-Eu-core**[2]**.** A email network of communication between institution members, which is generated using email data from a large European research institution. It consists of 1005 nodes and 25571 edges.
3) **polblogs**[1]**.** A network of hyperlinks between weblogs on US politics, which is recorded in 2005. It consists of 1490 nodes and 19090 edges.
4) **Cora**[3]**.** A citation network of scientific publications, which consists of maching learning papers that can be classified into seven classes. It consists of 2708 nodes and 5429 edges.

Datasets without ground-truth community labels:

1) **power**[4]**.** A network that represents the topology of the Western States Power Grid of the United States. It consists of 4941 nodes and 6594 edges.
2) **ca-GrQc**[2]**.** A collaboration network of scientific collaborations between authors with papers submitted to General Relativity and Quantum Cosmology category. It consists of 5242 nodes and 14496 edges.
3) **as-22july06**[4]**.** A symmetric snapshot of the structure of the Internet at the level of autonomous systems. It consists of 22963 nodes and 48436 edges.
4) **email-Enron**[4]**.** A email network that covers all the email communication within a dataset of around half million emails. It consists of 36692 nodes and 367662 edges.

Three commonly used evaluation measures are adopted for evaluating the quality of the discovered community structure, namely Normalized Mutual Information (NMI), F-score and Modularity (Q) [8]. The first two measures require the ground-truth community labels for evaluation purpose and the values are in the range between 0 and 1. The last measure Q ranges from -1 to 1. For all the three measures, a higher value indicates better performance.

4.2 Graph Partitioning Methods

We adopt the following four graph partitioning methods in our experiments:

1) **Louvain** [5]: It is a greedy community detection method that can reveal the hierarchical community structure.
2) **Spectral Clustering (SC)** [27]: It performs a spectral clustering of the node adjacency matrix into c clusters.
3) **Affinity Propagation (AP)** [9]: It is a fast clustering method based on similarities between pairs of data points.
4) **Nonnegative Matrix Factorization (NMF)** [6]: It obtains the new property representation by factorizing the node adjacency matrix into two nonnegative matrices.

Each of the above four graph partitioning methods is taken as the input "graph partitioning method `S`" of Algorithm 1, by which different variants of **EdMot** are obtained, namely EdMot-Louvain, EdMot-SC, EdMot-AP and EdMot-NMF. Similarly, as the conventional higher-order community detection methods, each of the above four graph partitioning methods takes as input the hypergraph of higher-order connectivity, i.e. the motif adjacency matrix $W_{\mathbf{M}}$, by which different variants of the higher-order community detection methods are obtained, namely Motif-Louvain, Motif-SC, Motif-AP and Motif-NMF. And the original graph partitioning method, the corresponding higher-order variant, and the corresponding **EdMot** variant are compared and analyzed. For instance, Louvain, Motif-Louvain and EdMot-Louvain are compared.

[1] http://www-personal.umich.edu/~mejn/netdata/
[2] http://snap.stanford.edu/data/
[3] http://linqs.cs.umd.edu/projects/projects/lbc/
[4] https://graph-tool.skewed.de/static/doc/collection.html

4.3 Comparison Results

Comparison results are reported from Table 1 to Table 5, where the scores are averaged over 20 runs for every method and the standard deviations are also reported. Table 1 to Table 4 report the results on datasets with ground-truth community labels while Table 5 is for two datasets without ground-truth community labels. Additionally, the average rank is also provided in the last column, which is computed by averaging the ranking positions of each method across the testing datasets.

Specifically, Table 1 reports the comparison results among Louvain, Motif-Louvain and EdMot-Louvain. As can be seen, the best scores are achieved by EdMot-Louvain in terms of NMI, F-score and modularity on polbooks, polblogs and Cora. On average, about 17% and 14% improvements (in terms of NMI) have been achieved by EdMot-Louvain over Louvain and Motif-Louvain respectively. This may be due to the utilization of edge enhancement and the original network structure, which plays an important role in avoiding hypergraph fragmentation and preserving the structural information of the network. On the email-Eu-core dataset, the proposed method performs worse than Motif-Louvain in terms of NMI. The reason is that this dataset has a relatively denser linkage structure (i.e., 1005 nodes and 25571 edges) and the hypergraph fragmentation issue does not appear on this dataset. That is, the motif-based hypergraph is a single connected component. While the polblogs and Cora datasets suffer from the hypergraph fragmentation issue, which accounts for the better performance achieved by the proposed method. As for the polbooks dataset, even though the hypergraph fragmentation does not appear, the proposed method still performs well. Similar analysis can be made in Table 2 to Table 5.

In summary, the proposed method can perform better than the original graph partitioning methods and the traditional higher-order methods with hypergraph fragmentation issue. What's more, it may also improve the performance of some state-of-the-art graph partitioning methods despite there is no hypergraph fragmentation issue.

4.4 Parameter Analysis

In this section, parameter analysis is conducted to investigate the effect of the parameter K on the performance of our **EdMot** method. Due to the space limit, we will take the Louvain method as the input graph partitioning method. However, similar analysis can be conducted by using any other graph partitioning methods.

Table 1: Comparison results on the Louvain method. The best result in each measure is highlighted in bold.

		polbooks	email-Eu-core	polblogs	Cora	Avg.rank
NMI	Louvain	0.4142±0.0000	0.5642±0.0000	0.2684±0.0000	0.3996±0.0000	2.75
	Motif-Louvain	0.4818±0.0000	**0.6350**±0.0000	0.2513±0.0000	0.4043±0.0000	2.00
	EdMot-Louvain	**0.4981**±0.0000	0.6098±0.0000	**0.3464**±0.0000	**0.4088**±0.0000	1.25
F-score	Louvain	0.4954±0.0000	0.5430±0.0000	0.6117±0.0000	0.0486±0.0000	2.50
	Motif-Louvain	0.6340±0.0000	0.5149±0.0000	0.5858±0.0000	0.0517±0.0000	2.50
	EdMot-Louvain	**0.6462**±0.0000	**0.5735**±0.0000	**0.7613**±0.0000	**0.0690**±0.0000	1.00
Modularity	Louvain	0.4833±0.0000	0.3933±0.0000	0.4272±0.0000	0.5439±0.0000	2.50
	Motif-Louvain	0.5058±0.0000	0.4024±0.0000	0.4221±0.0000	0.4216±0.0000	2.50
	EdMot-Louvain	**0.5092**±0.0000	**0.4085**±0.0000	**0.4307**±0.0000	**0.5783**±0.0000	1.00

Table 2: Comparison results on the SC method. The best result in each measure is highlighted in bold.

		polbooks	email-Eu-core	polblogs	Cora	Avg.rank
NMI	SC	0.5422±0.0000	**0.7049**±0.0034	0.0016±0.0000	0.3953±0.0000	2.25
	Motif-SC	0.5458±0.0102	0.6509±0.0051	0.0030±0.0016	0.0913±0.0366	2.50
	EdMot-SC	**0.5675**±0.0000	0.6913±0.0022	**0.3975**±0.0000	**0.4226**±0.0038	1.25
F-score	SC	0.8216±0.0000	**0.5533**±0.0076	0.7580±0.0000	0.5392±0.0000	2.00
	Motif-SC	0.8298±0.0059	0.4896±0.0076	0.6541±0.0634	0.3562±0.0185	2.75
	EdMot-SC	**0.8416**±0.0000	0.5419±0.0027	**0.8112**±0.0000	**0.5728**±0.0265	1.25
Modularity	SC	0.5015±0.0000	**0.2598**±0.0029	0.0007±0.0000	0.6599±0.0000	2.25
	Motif-SC	0.5041±0.0032	0.2410±0.0033	0.0009±0.0003	0.3674±0.0569	2.50
	EdMot-SC	**0.5099**±0.0000	0.2520±0.0014	**0.4309**±0.0000	**0.7108**±0.0103	1.25

Table 3: Comparison results on the AP method. The best result in each measure is highlighted in bold.

		polbooks	email-Eu-core	polblogs	Cora	Avg.rank
NMI	AP	0.3541±0.0138	0.4519±0.0149	0.1238±0.0024	0.3856±0.0015	3.00
	Motif-AP	0.3974±0.0077	**0.6318**±0.0011	**0.1772**±0.0003	0.3962±0.0010	1.50
	EdMot-AP	**0.4029**±0.0044	0.5909±0.0062	0.1601±0.0028	**0.3973**±0.0021	1.50
F-score	AP	0.2510±0.0116	0.2075±0.0144	0.0983±0.0026	0.0666±0.0016	2.50
	Motif-AP	0.2748±0.0211	0.2219±0.0014	**0.1046**±0.0006	0.0481±0.0009	2.00
	EdMot-AP	**0.2873**±0.0129	**0.3185**±0.0053	0.0980±0.0117	**0.0750**±0.0046	1.50
Modularity	AP	0.2214±0.0151	0.0749±0.0065	0.1045±0.0026	0.4723±0.0026	2.75
	Motif-AP	0.2977±0.0114	0.1712±0.0010	**0.2251**±0.0009	0.3821±0.0024	2.00
	EdMot-AP	**0.3069**±0.0159	**0.2187**±0.0029	0.1148±0.0098	**0.4770**±0.0321	1.25

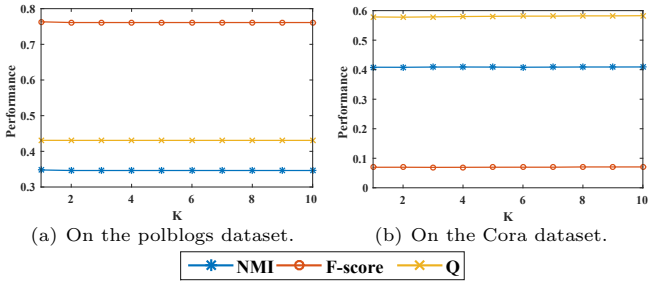
Table 4: Comparison results on the NMF method. The best result in each measure is highlighted in bold.

		polbooks	email-Eu-core	polblogs	Cora	Avg.rank
NMI	NMF	0.4528±0.0086	0.7090±0.0063	0.4005±0.0000	0.2723±0.0071	2.25
	Motif-NMF	0.4809±0.0302	0.6594±0.0060	0.2687±0.0100	0.1010±0.0058	2.75
	EdMot-NMF	**0.5568**±0.0042	**0.7113**±0.0042	**0.4040**±0.0012	**0.2821**±0.0028	1.00
F-score	NMF	0.7750±0.0101	0.4756±0.0016	0.7644±0.0000	0.7644±0.0016	2.00
	Motif-NMF	0.7326±0.0449	**0.7644**±0.0119	0.7644±0.0050	0.7644±0.0094	1.75
	EdMot-NMF	**0.8382**±0.0024	0.5811±0.0019	**0.8145**±0.0003	0.4854±0.0033	1.50
Modularity	NMF	0.4484±0.0060	0.2663±0.0055	0.4305±0.0000	**0.6499**±0.0010	2.00
	Motif-NMF	0.4808±0.0078	0.2729±0.0072	0.4219±0.0005	0.4407±0.0105	2.25
	EdMot-NMF	**0.4999**±0.0019	**0.2740**±0.0087	0.4305±0.0000	0.6216±0.0099	1.25

The effect of K on the performance of EdMot-Louvain is shown in Figure 3 and Figure 4. Specifically, the effect of K on polblogs and Cora is presented in Figure 3. Since the motif-based hypergraphs of polbooks and email-Eu-core contain only one connected component, the discussion for these two datasets is omitted here. As can be seen, the scores of the evaluation measures hardly change as K increases. This is because the motif-based hypergraph often contains

Table 5: Comparison results in terms of modularity on the two datasets without ground-truth community labels.

Method	power	ca-GrQc	Avg.rank
Louvain	0.5248±0.0000	0.6863±0.0000	3.00
Motif-Louvain	0.8053±0.0000	0.7806±0.0000	2.00
EdMot-Louvain	**0.9983**±0.0000	**0.9988**±0.0007	1.00
SC	0.8545±0.0000	0.1426±0.0037	2.50
Motif-SC	0.8159±0.0034	0.6725±0.0007	2.50
EdMot-SC	**0.8607**±0.0023	**0.7279**±0.0027	1.00
AP	0.4907±0.0020	0.4220±0.0000	2.00
Motif-AP	0.1799±0.0005	0.4169±0.0035	3.00
EdMot-AP	**0.8551**±0.0042	**0.5483**±0.0064	1.00
NMF	0.4922±0.0015	0.5433±0.0051	3.00
Motif-NMF	0.6336±0.0191	0.5704±0.0171	2.00
EdMot-NMF	**0.9406**±0.0151	**0.8920**±0.0058	1.00



(a) On the polblogs dataset. (b) On the Cora dataset.

Figure 3: Parameter analysis: Effect of K on two datasets with ground-truth community labels.

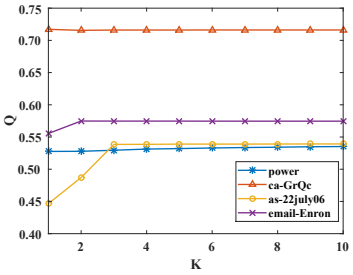


Figure 4: Parameter analysis: Effect of K on four datasets without ground-truth community labels.

a largest connected component that may consist of nodes from different communities and a large number of peripheral nodes that are isolated (see Figure 1(b) for illustration). Therefore, it is beneficial to partition the largest connected component in the hypergraph ($K = 1$) into modules. Similar phenomenon can also be observed from the power and ca-GrQc datasets as shown in Figure 4. As for the as-22july06 ($n = 22963$, $m = 48436$) and the email-Enron ($n = 36692$, $m = 367662$) datasets, which are relatively larger than other datasets, the best performance is achieved when $K = 3$ and $K = 2$ respectively. Thus, we set $K = 3$ for the as-22july06 dataset, $K = 2$ for the email-Enron dataset and $K = 1$ for the other datasets in the experiments.

5 RELATED WORK

5.1 Lower-Order Community Detection

Lower-order community detection methods discover communities by mainly leveraging the lower-order connectivity patterns of the network at the level of individual nodes and edges. For example, the Louvain method was proposed to reveal the hierarchical community structure by heuristically optimizing modularity [5]. The Nonnegative Matrix Factorization (NMF) method factorizes the node adjacency matrix into two nonnegative matrices and yields a new parts-based data representation [6, 7]. From the perspective of clustering nodes in the network into clusters, the Affinity Propagation (AP) clustering method was proposed to cluster nodes based on pair-wise similarities [9]. Similarly, the Spectral Clustering (SC) method was proposed to cluster nodes using eigenvectors of matrix [27]. Besides, a generative model termed Stochastic BlockModel (SBM) was proposed to detect communities by fitting blockmodels to the network data [17, 28]. And the permanence based method was also proposed to provide a more fine-grained view of the modular structure of the network [8]. In addition, the label propagation based methods were developed, which possess some advantages such as the simplicity and nearly linear time complexity [30].

5.2 Motif-based Higher-Order Community Detection

Network motifs are defined as patterns of interconnections occurring in networks at numbers that are significantly higher than those in the random networks [3, 22]. As the building blocks of the network, motifs are widely applied to unravel the design principles of gene regulation networks [32] and the underlying mechanisms of social networks [11, 35].

Different from the lower-order community detection methods, higher-order community detection methods discover communities by leveraging the higher-order connectivity patterns of the network at the level of small network subgraphs, e.g., motifs. For example, motif was used to define communities by extending the mathematical expression of Newman Girvan modularity in [2]. A graph sparsification principle based on graph motifs was proposed so as to improve the efficiency and quality for graph partitioning [37]. The motif conductance based framework was also proposed to reveal higher-order organization of complex networks [3]. Besides, a highly effective heuristic method was proposed based on motifs, where a random walk interpretation of the graph reweighing scheme was developed [34]. In addition, motifs have been leveraged in local higher-order graph partitioning [36].

Despite the success in preserving the building blocks, the existing higher-order community detection methods may violate the original lower-order structure of the network. Specifically, the motif-based hypergraph may consist of a large number of connected components with various sizes and isolated nodes having no connections to any other nodes. In this case, they may suffer seriously from the hypergraph fragmentation issue. To address this issue, we propose an

edge enhancement approach for motif-aware community detection (**EdMot**), which can not only leverage higher-order connections of the network but also overcome the hypergraph fragmentation issue.

6 CONCLUSION

In this paper, we for the first time propose a novel motif-aware community detection method termed **EdMot** for addressing the hypergraph fragmentation issue. Different from the existing higher-order community detection methods that directly operate on the possibly fragmented hypergraph, **EdMot** partitions the top K largest connected components in the hypergraph into modules. And then, an edge enhancement approach is designed for enhancing the connectivity structure of the original network as follows. 1) First, a new edge set is constructed to derive a clique from each module. 2) Based on the new edge set, the original connectivity structure of the input network is enhanced to generate a rewired network, whereby the motif-based higher-order structure is leveraged and the hypergraph fragmentation issue is well addressed. After the edge enhancement, the rewired network is partitioned to obtain the higher-order community structure. Extensive experiments have been conducted to show the effectiveness of the proposed method.

ACKNOWLEDGMENTS

This work was supported by NSFC (61876193), National Key Research and Development Program of China (2016YFB1001003), Guangdong Natural Science Funds for Distinguished Young Scholar (2016A030306014), and Key Areas Research and Development Program of Guangdong (2018B010109007).

REFERENCES

[1] Charu C. Aggarwal and Haixun Wang. 2010. *Managing and Mining Graph Data.*
[2] 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.
[3] Austin R Benson, David F Gleich, and Jure Leskovec. 2016. Higher-order organization of complex networks. *Science* 353, 6295 (2016), 163–166.
[4] Jonathan W. Berry, Luke K. Fostvedt, Daniel J. Nordman, Cynthia A. Phillips, C. Seshadhri, and Alyson G. Wilson. 2014. Why do simple algorithms for triangle enumeration work in the real world?
[5] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of community hierarchies in large networks. *J Stat Mech* abs/0803.0476 (2008).
[6] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. 2011. Graph regularized nonnegative matrix factorization for data representation. *IEEE TPAMI* 33, 8 (2011), 1548–1560.
[7] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. 2008. Non-negative matrix factorization on manifold. In *ICDM*. 63–72.
[8] Tanmoy Chakraborty, Sriram Srinivasan, Niloy Ganguly, Animesh Mukherjee, and Sanjukta Bhowmick. 2014. On the permanence of vertices in network communities. In *KDD*. 1396–1405.
[9] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* 315, 5814 (2007), 972–976.
[10] Lifang He, Chun-Ta Lu, Jiaqi Ma, Jianping Cao, Linlin Shen, and Philip S Yu. 2016. Joint community and structural hole spanner detection via harmonic modularity. In *KDD*. 875–884.
[11] Paul W Holland and Samuel Leinhardt. 1977. A dynamic model for social networks. *Journal of Mathematical Sociology* 5, 1 (1977), 5–20.
[12] Paul W Holland and Samuel Leinhardt. 1977. A method for detecting structure in sociometric data. In *Social Networks*. Elsevier, 411–432.
[13] Ling Huang, Chang-Dong Wang, and Hong-Yang Chao. 2018. A Harmonic Motif Modularity Approach for Multi-layer Network Community Detection. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*. 1043–1048.
[14] Ling Huang, Chang-Dong Wang, and Hong-Yang Chao. 2018. Overlapping Community Detection in Multi-view Brain Network. In *BIBM*. 655–658.
[15] Ling Huang, Chang-Dong Wang, and Hong-Yang Chao. 2019. Higher-Order Multi-layer Community Detection. In *AAAI*.
[16] Marcus Kaiser. 2008. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New Journal of Physics* 10, 8 (2008), 083042.
[17] Brian Karrer and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical review E* 83, 1 (2011), 016107.
[18] Matthieu Latapy. 2008. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical computer science* 407, 1-3 (2008), 458–473.
[19] Juan-Hui Li, Chang-Dong Wang, Pei-Zhen Li, and Jian-Huang Lai. 2018. Discriminative metric learning for multi-view graph partitioning. *Pattern Recognition* 75 (2018), 199–213.
[20] Pei-Zhen Li, Yue-Xin Cai, Chang-Dong Wang, Mao-Jin Liang, and Yi-Qing Zheng. 2018. Higher-order Brain Network Analysis for Auditory Disease. *Neural Processing Letters* (2018).
[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] 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.
[23] Mark EJ Newman. 2004. Detecting community structure in networks. *The European Physical Journal B* 38, 2 (2004), 321–330.
[24] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.
[25] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113.
[26] Mark EJ Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E* 68, 3 (2003), 036122.
[27] Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *NIPS*. 849–856.
[28] Tiago P Peixoto. 2014. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E* 89, 1 (2014), 012804.
[29] 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 TKDD* 10, 3 (2016), 22.
[30] 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.
[31] Junming Shao, Zhichao Han, Qinli Yang, and Tao Zhou. 2015. Community detection based on distance dynamics. In *KDD*. 1075–1084.
[32] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics* 31, 1 (2002), 64.
[33] Steven H Strogatz. 2001. Exploring complex networks. *Nature* 410, 6825 (2001), 268.
[34] Charalampos E Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. 2017. Scalable motif-aware graph clustering. In *WWW*. 1451–1460.
[35] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press.
[36] Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. 2017. Local higher-order graph clustering. In *KDD*. 555–564.
[37] Peixiang Zhao. 2015. gSparsify: Graph Motif Based Sparsification for Graph Clustering. In *CIKM*. 373–382.