IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 May 2023, accepted 7 June 2023, date of publication 19 June 2023, date of current version 22 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287214

RESEARCH ARTICLE

Identifying Multiple Propagation Sources With Motif-Based Graph Convolutional Networks for Social Networks

KAIJUN YANG, QING BAO, AND HONGJUN QIU
School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China

Corresponding author: Qing Bao (qbao@hdu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61806061, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ19F030011.

ABSTRACT Identifying the sources of propagation in social networks, such as the misinformation propagation, is one of the key issues recently. Most existing studies assume the underlying propagation model is known, which is difficult to obtain in practice. Recent efforts have been devoted to detect multiple sources in real-world situations, and the social influence of neighbors in the propagation is assumed to be identical. However, this assumption will result in inaccurate results as the infection state of a node is determined by its critical neighbors. In this paper, we fill this gap by capturing social influence of neighbors with structural properties in social networks. For instance, opinions are more likely to spread via closely connected friends within small groups. Here we propose a Motif-based Graph Convolutional Networks for Source Identification (MGCNSI) framework based on the GCN-based source identification approach. Specifically, different network motifs are used to capture different types of structural properties. Then each motif extracts the critical neighbors of a particular type, and a motif-based graph convolutional layer is constructed to aggregate critical neighbors for that motif. To adapt to underlying propagation mechanisms, an attention mechanism for aggregation is designed to automatically assign higher weights to more informative motifs. The empirical results demonstrate that MGCNSI outperforms several benchmark methods on both synthetic and real-world networks. The advantage is most obvious for networks with denser node neighborhoods, where MGCNSI can select critical neighbors from the larger neighbor sets. How the motifs can capture the social influence and the underlying critical paths of propagation is also illustrated.

INDEX TERMS Information propagation, multiple source identification, motif, graph convolutional networks.

I. INTRODUCTION

The emergence and development of communication technologies and online social networks have dramatically changed our lifestyle. This not only makes our daily life more convenient, but also makes us vulnerable to various network risks. For instance, the connected users on social networking sites like Facebook and Twitter can share individual opinions and information, which may contain incorrect information. As a consequence, rumors can be shared and forwarded rapidly on those social networking sites [1], [2], [3]. To prevent and control a harmful spread on networks, it is critical to identify the possible origins accurately. If the origins could be identified in the early stage, intervention and control strategies can be developed to curtail the spread, and diminish potential damages. Consequently, the problem of source identification on networks has attracted lots of attention in the past few years [4], [5].

In recent years, extensive approaches have been proposed to solve the source identification problem on networks. Given a network and the observed configurations, the goal is to find the most probable propagation sources on the network. Early

The associate editor coordinating the review of this manuscript and approving it for publication was Zhipeng Cai.

studies mainly focused on spreading dynamics on tree-like networks [6], [7], [8]. Later on, the problem was generalized to general complex networks [9], [10], and the methods to identify multiple sources were also proposed [11], [12], [13]. Recently, to adapt to specific situations, some researchers developed methods on time-varying networks [14], [15], [16] and multilayer networks [17].

One of the main challenges of source identification lies in the stochastic nature of the spreading dynamics. Until now, most studies assumed the underlying propagation model to be known in advance, and various models have been used to simulate the spreading dynamics, such as the Independent Cascade model [18], the Susceptible-Infectious model [19], the Susceptible-Infectious-Recovered model [20], and etc. To reflect different real-world situations, other types of spreading dynamics such as the Susceptible-Exposed-Infectious-Recovered (SEIR) model were introduced to the source identification problem [21], [22], [23]. In practice, however, identifying the actual propagation model requires detailed domain knowledge, which is often difficult to obtain.

Until recently, some recent efforts have been devoted to detect multiple sources without knowing the specific propagation model [24], [25], [26], [27]. The only assumption is that sources are more likely to be surrounded by a larger proportion of infected nodes. Among them, the LPSI [24] method was firstly proposed to classify the nodes into sources or not based on the label propagation algorithm, and the GCNSI [25] method exploited the graph convolutional networks to enhance the accuracy of LPSI method. Other approaches utilized the invertible graph diffusion models [26] or variational autoencoders [27] to solve the inverse problem of forward diffusion estimation. Nonetheless, as the specific propagation model can not be obtained, the above-mentioned methods have the same assumption that the influence of each neighbor node is assumed to be identical in the propagation, which will result in inaccurate results.

In this paper, we investigate the source identification problem with unknown propagation mechanism, and focus on capturing and incorporating the social influence of neighbors for source identification. To the best of our knowledge, this is the first work to tackle this problem. In particular, the mesoscopic structural properties of networks are utilized to capture the social influence. Existing studies have shown that rich social network structural properties contain useful information of social influence [28], [29], [30]. For example, opinions are more likely to spread via closely connected friends within small groups. The groups of densely connected nodes are called communities, which have denser intra-group connections than the inter-group ones [29], [31], [32]. For more complex spreading dynamics, social affirmation from multiple neighbors is required. As another example, we may get convinced when we are a social group of three individuals, and our two neighbors are both adopters. In those cases, a hyperedge (also called hyperlink) with more than two nodes can be adopted to model such social groups [30], [33]. The importance of weak ties acting as ''local bridges'' in facilitating effective information diffusion in social networks, has also long been recognized [28], [34], [35]. For instance, information provided by friends in other groups carries less redundancy. For the past decades, researchers working on information diffusion modeling have found that the observed information cascades can be explained by various mesoscopic structures such as communities, hyperedges, and etc [32], [33], [36], [37], [38].

We here propose a novel Motif-based Graph Convolutional Networks for Source Identification (MGCNSI) framework. The Graph Convolutional Network based approach is adopted to identify source nodes. The node embeddings are obtained via propagation of neighbors' features, but here the difference in the social influence of neighbors needs to be considered. To achieve this goal, some challenges need to be solved:

Firstly, how to effectively capture the patterns of common structural properties? Here MGCNSI utilizes network motifs, which are often used to represent over-represented patterns of subgraphs. As one of the most common higher-order structures, network motifs have been extensively adopted to capture the structural and functional properties of network data [39], [40], [41], [42]. In particular, networks in the same domain are composed of similar motifs, whereas networks from different domains have significantly different motif frequencies [39], [40]. In social networks, similar patterns such as friends in small close groups, or different groups can also be depicted by certain motifs based on their specific functions in local network structures [43], [44].

Secondly, how to capture different forms of social influence? Given a set of network motifs, multiple motif networks can be constructed, where each motif network selects sets of critical neighbors of a particular type of social influence, such as the influence of friends in the same social group. Different motif networks extract different sets of critical neighbors for each node. Based on that, a motif-based graph convolutional layer can be designed and applied to each motif network to aggregate the critical neighbors of a particular type.

Furthermore, different forms of social influence are not equally important. An attention mechanism for aggregation is designed to automatically assign higher weights to more informative motifs. Hence, the critical neighbors of the underlying propagation mechanism can be depicted.

The main contributions of this paper are summarized as follows:

1) Different forms of social influence in the propagation, captured with structural properties in the social networks, are incorporated to identify propagation sources when the underlying propagation model is unknown.
2) The Motif-based Graph Convolutional Networks for Source Identification (MGCNSI) framework is proposed based on the GCN-based source identification approach. The critical neighbors with different forms of social influence are captured with network motifs and aggregated through multiple motif-based GCN layers. Also, an attention mechanism for aggregation is

designed to automatically assign higher weights to the informative motifs.

3) A series of simulations on both synthetic and real-world networks are conducted to evaluate the effectiveness of the proposed method. Also, the results demonstrate how the motifs can support the analysis for social influence and the underlying critical paths of propagation.

The remainder of this paper is organized as follows. Section II presents some related work. Section III introduces the definition of multi-source identification with the underlying propagation model unknown. Section IV shows the MGCNSI framework in detail and Experimental results are reported in Section V. Finally, Section VI concludes this paper and provides pointers for future work.

II. RELATED WORK

A. SOURCE IDENTIFICATION

In the past decades, researchers have proposed a variety of methods to identify propagation sources. The early studies mainly focused on spreading dynamics on tree-like networks and proposed several types of centrality scores [6], [7], [8]. Similar to the centrality scores to identify influential spreaders for reaching a maximum spreading ability [45], [46], [47], [48], [49], here the scores were introduced to identify candidate sources by maximizing the likelihood of the obtained traces. For example, Shah and Zaman proposed a rumor centrality score for the Susceptible-Infected (SI) model and proved that the node with the largest rumor centrality score can maximize the likelihood of the observed data [6], [7]. Zhu and Ying further proposed Jordan centrality scores for the Susceptible-Infectious-Recovered (SIR) model, and obtained candidate propagation sources by selecting the nodes with the highest Jordan centrality scores [8].

Later on, the source identification problem was generalized from tree-like networks to general networks [9], [10]. For instance, Kazemitabar [10] proposed to compute the Bayes optimal solution to identify the candidate sources from a complete snapshot of the infected nodes in general networks. Also, the methods to identify multiple sources were developed [11], [12], [13]. For instance, Prakash et al. [11] proposed to search the multiple information sources based on the Minimum Description Length principle under the SI model. Zang et al. [12] further introduced a reverse propagation model to detect the recovered and unobserved infected nodes, so as to identify multiple sources under the SIR model. Considering the dynamics of the networks and multi-layer nature of human interactions, methods for time-varying networks [14], [15], [16] and multilayer networks [17] were developed recently to adapt to those specific situations. In terms of observations, while most studies obtained a complete snapshot of infected nodes, some attempted to inject multiple sensors in networks to obtain the state changes of the sensor nodes and the infection time [22], [50], [51]. The inactivated sensor nodes were also considered recently [52]. This paper studies the multiple source identification problem on general networks given complete steady-state snapshots, which is the most widely used settings in recent years.

Another challenge of the source identification problem lies in the underlying spreading dynamics of the observed configuration. Most existing studies assumed that the underlying propagation model is known in advance, and propagation models like the Independent Cascade model [18], the Susceptible-Infectious model [19], and the Susceptible-Infectious-Recovered model [20] have been adopted to simulate the propagation dynamics [53]. Recently, the Susceptible-Exposed-Infectious-Recovered (SEIR) model was introduced to include the latent exposed period, i.e., the period from the time of infection to the time of becoming infectious [21]. The incubation period of the infectious diseases, i.e., the time period between when an individual becomes infected and when the symptoms start, was considered to identify sources of asymptomatic spread [22]. To detect sources of rumor dissemination, Chen et al. added ''fact checkers'' to define the Denied state when an individual refutes the rumor [23]. The assumption of the underlying propagation mechanism has an important impact on the performance of source identification, however, it is difficult to obtain the underlying propagation model in practice.

The LPSI [24] method, proposed based on the label propagation algorithm, was the first work to identify sources without any information of the underlying propagation model. The only assumption is that the sources are more likely to be surrounded by a larger proportion of infected nodes. Later on, several deep learning approaches have been proposed. Among them, GCNSI [25] method exploited the Graph Convolutional Network approach to enhance the accuracy of the LPSI method. As source identification is the inverse of the diffusion process on graphs, Wang et al. [26] built an invertible graph diffusion model named invertible graph residual net to solve the inverse problem. Ling et al. [27] proposed a probabilistic model that leveraged graph generative models to capture a generative prior and conditional probability of forward diffusion estimation, so that the probable sources could be inferred via a variational inference-based method. Jiang et al. [54] realized rumor source identification, rumor news detection, and popularity prediction into a joint learning framework, and identified the candidate rumor source clusters to reduce the search space before applying a multilayer perceptron module to predict source nodes. However, the influence of each neighbor node is assumed to be identical in those approaches, which is often not the case in real situations. In this paper, we proposed that different forms of social influence in the propagation can be captured with structural properties in the social network, and incorporated it to improve the accuracy of source identification.

B. MESOSCOPIC STRUCTURE PROPERTIES IN SOCIAL NETWORKS

Properties of social networks have been analyzed in the past decades for various applications such as diffusion predic-

tion [33], [36], [55], influential spreaders identification [47], [48], [49] and etc. In particular, mesoscopic structures of social networks can indicate the social influence of different neighbors and researchers working on information diffusion modeling have found that the observed information cascades with different propagation mechanisms can be explained by such mesoscopic structures [32], [33], [36], [37], [38]. For instance, the observed information cascades can be explained by community structure [32], [36], [56], [57]. As closely connected friends within the same community have similar opinions in real-world situations, Barbieri et al. [32], [56] grouped the individuals into communities and modeled the influence among the communities. In addition, Bao et al. [36] proposed a component-based diffusion model which assumes that the influence of the neighboring nodes to a given node is not exerted individually but by connected components. A community detection algorithm was applied to the neighborhood of each node to identify the underlying components. The components and information diffusion can also be learnt jointly in [57]. Recent studies considered more complex spreading dynamics where social affirmation from all individuals in a social group is needed, and a hyperedge with more than two nodes was adopted to model such social groups [30]. Based on the traditional Susceptible-Infectious model, Iacopini et al. [33] introduced a higher-order model of social contagion, in which the contagion can occur through interactions in groups of different sizes. Jhun et al. [58] considered the higher-order contagion model based on the Susceptible-Infectious-Susceptible model. The higher-order contagion model on temporal networks [37] and multilayer networks [38] have also been developed.

C. DEEP LEARNING ON GRAPH DATA

Graph Neural Networks (GNN) have been widely used on graph data recently [59]. The idea of Graph Neural Networks (GNN) is inspired by the recent success of neural networks for Euclidean data, and it extends traditional deep learning methods for non-Euclidean data represented by graphs with complex relationships and interdependency between objects. Inspired by traditional convolutional neural networks, Bruna et al. [60] developed a graph convolution propagation rule based on the spectral graph theory, which is the earliest work of spectral-based convolutional graph neural networks. Since then, there have been increasing extensions [61], [62], [63], [64]. For example, Kipf et al. [64] presented a scalable approach based on an efficient variant of convolutional neural networks via a localized first-order approximation of spectral graph convolutions. Later on, many spatial-based convolutional graph neural networks emerged [65], [66], [67]. Recently, while some researchers developed advanced algorithms such as GNN acceleration algorithms [68], some focused on real-world applications.

Recently, an increasing number of successful applications of Graph Neural Networks have emerged, as graph-structured data are ubiquitous and GNNs can effectively the hidden patterns of non-Euclidean data. For social networks, common tasks like node classification [64], [69] and graph classification [70], and a variety of real-world problems such as community detection [29], [71], social influence prediction [72], information diffusion prediction [55], [73], and source identification [25] can be solved by GNNs.

TABLE 1. Summarization of notations.

Notation	Definition
$\mathcal{G}$	The social network.
$\mathcal{V}$	Set of nodes in $\mathcal{G}$.
$\mathcal{E}$	Set of edges in $\mathcal{G}$.
N	Number of nodes in $\mathcal{G}$, i.e., $N = \lvert\mathcal{V}\rvert$.
$\mathcal{Y} = [Y_1, \cdots, Y_N]^\top$	Infection state of the network, Y_i represents the infection state of node i.
S_p	Predicted source set.
S_g	Ground-truth source set.
$\mathbf{Y}' = [\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4]$	Expanded infection state.
C	Number of motifs.
$\mathcal{M} = \{M_1, \cdots, M_C\}$	Set of network motifs.
$\mathcal{A} = \{\mathbf{A}_1, \cdots, \mathbf{A}_C\}$	Set of adjacency matrices for motif networks of network $\mathcal{G}$.
$\mathcal{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_C\}$	Set of embeddings for $\mathcal{A}$.
$\mathbf{X}_f$	Final node embeddings for $\mathcal{G}$.
$\boldsymbol{\alpha}_i$	The nodes' attention coefficients for motif M_i. α_{ij} represents the attention coefficient of node j for motif M_i.
$\mathbf{W}_b, \mathbf{W}_e, \mathbf{W}_c^l, \mathbf{W}_f, \mathbf{W}_t, \mathbf{W}_\alpha^i$	Learnable weight matrices of MGCNSI.
$\mathbf{b}_c^l, \mathbf{b}_f$	Learnable bias vectors of MGCNSI.
λ	Weight coefficient in L2 regularization.
H	Hidden size, i.e., the dimension of the hidden layer in GCN.
$\mathbf{y}'$	Output of MGCNSI.
$\mathbf{y}$	Ground-truth of the source nodes.

III. PRELIMINARIES

In this section, we will first introduce the objectives of multi-source identification problem and then describe the related techniques required for this paper. Frequently used notations for this paper are shown in Table 1.

A. PROBLEM STATEMENT

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{Y}\}$ be an undirected social network, where $\mathcal{V}$ is a set of N nodes, $\mathcal{E} = \{(i,j) | i,j \in \mathcal{V}\}$ is the set of edges. $\mathcal{Y} = [Y_1, \cdots, Y_N]^\top$ is the infection state of the network, where $Y_i = 1$ indicates node i is infected in the propagation, and $Y_i = -1$ indicates that node i is not infected in the propagation.

The goal of our multiple sources identification problem is to minimize the difference between the predicted set of source nodes S_p and the ground-truth set of source nodes S_g. Specifically, our goal is to find a function $f : \mathcal{V} \to \{1, 0\}$ that maximizes (1):

$$\frac{|S_g \cap S_p|}{|S_g \cup S_p|} \quad with \;\; S_p = \{x \in \mathcal{V} \mid f(x) = 1\}. \tag{1}$$

B. PROPAGATION MODELS

In the field of information propagation, two kinds of propagation models are usually used for simulation, namely the infection models and the influence models [74]. There are two typical iterative influence models, namely the Independent Cascade (IC) model [18] and the Linear Threshold (LT) model [75]. In each iteration of the IC model, each newly activated node has one chance to activate its neighboring nodes, if a neighbor is not activated by the node, the neighbor will not be activated by this node again. For the LT model, each node will accumulate the influence of its activated neighbors. In each iteration, a node will be activated if the total influence exceeds the threshold. The simple propagation models (such as the IC model) usually consider one-to-one infection, whereas the complex propagation models (such as the LT model) consider the collective influence of multiple neighbors.

On the other hand, for the infection models, a node may typically have three states: S (Susceptible), I (Infected), and R (Recovered). Common infection models include the Susceptible-Infectious (SI) model, Susceptible-Infectious-Recovered (SIR) model and Susceptible-Infectious-Susceptible (SIS) model [19], [20]. The SI model has only two states, and a node has a probability to change from Susceptible state to Infected state in each iteration. The SIR model introduces the Recovered state based on the SI model. In each iteration, an infected node has a probability to change from the Infected state to the Recovered state, and the node will never be infected again. In the SIS model, a node can be infected multiple times.

In addition to the above basic propagation mechanisms, the social attributes in social networks will also affect the information propagation [32], [33], [36], [58]. The studies showed that the spread of information in social networks is related to group behavior, and the following situations usually exist: (i) Opinions are more likely to spread via closely connected friends within small groups. That is, the probability of propagation between two nodes within the same community is high, whereas the probability between two nodes of different communities is low [32], [36]; (ii) We may get convinced when we are in a social group, and other members of the group are adopters. That is, a node can be affected not only by its direct neighbors but also by its group, with different probabilities [33], [58].

C. SOURCE IDENTIFICATION BASED ON LABEL PROPAGATION

Wang et al. [24] proposed Label Propagation based Source Identification (LPSI) based on the label propagation algorithm. This is the first work to identify sources without any information of the underlying propagation model. It can be used to generate input features for more advanced models. In LPSI, each node is assigned a label. The initial label of each node i is the infection state Y_i. The node labels are propagated iteratively and it will finally converge. The final label of each node indicates the probability of the node being a propagation source. In particular, The iteration formulation of LPSI is as follows:

$$g_i^{t+1} = \mu \sum_{j:j\in\mathcal{N}(i)} S_{ij} g_j^t + (1-\mu) Y_i, \tag{2}$$

where $\mathbf{g}^t$ represents the vector of node labels in the t^{th} iteration, and g_i^t represents the value for node i. $\mathcal{N}(i)$ is the set of neighbors of node i in the network. $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, $\mathbf{A}$ is the adjacency matrix of $\mathcal{G}$, $\mathbf{D}$ is the Laplacian matrix of $\mathcal{G}$, Y_i is the infection state of node i, and μ controls the influence of neighbors. It was proved that the convergent node labels are given as:

$$\mathbf{g}^* = (1-\mu)(\mathbf{I} - \mu\mathbf{S})^{-1}\mathcal{Y}, \tag{3}$$

where $\mathcal{Y}$ is the infection state of the network, $\mathbf{g}^*$ is the vector of final node labels when convergence is achieved. LPSI has two versions, namely the iterative version (2) and the convergent version (3).

As shown in (2), the influence of each neighbor is equal in LPSI. This is due to the fact that social influence of neighbors can not be easily obtained as the underlying propagation model is not known. However, in real situations of information propagation in social networks, it is the critical neighbors that determine the infection state of a node. In terms of source identification, the assumption of equal influence in LPSI and other source identification approaches will result in inaccurate results. In this paper, different forms of social influence in the propagation are captured with structural properties in the social network. The detailed method and experimental results will be shown in the next two sections.

D. SOURCE IDENTIFICATION BASED ON GRAPH CONVOLUTIONAL NETWORKS

Traditional convolutional neural networks (CNNs) are widely used in the field of computer vision and have achieved great success. Graph convolution neural network (GCN) [60] extends CNN to graph data to capture node features of a graph effectively. Early versions of GCN were very time-consuming. Hammond [61] introduced the Chebyshev polynomial to speed up the calculation. Kipf's work further simplified GCN [64] and we adopt this version. Specifically, given the initial node features as input, graph convolutional operations are performed for several layers, and the final node features can be obtained from the output of the last layer. The layer-wise graph convolutional operation is defined as:

$$\mathbf{H}^{l+1} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^l\mathbf{W}^l), \tag{4}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{A}$ is the adjacency matrix, $\mathbf{I}$ is the identity matrix, $\tilde{\mathbf{D}}$ is the Laplacian matrix of $\tilde{\mathbf{A}}$, $\mathbf{H}^l$ is the node features of l^{th} layer, $\mathbf{W}^l$ is the trainable weight matrix of l^{th} layer, and $\sigma(\cdot)$ is the nonlinear activation function. $\mathbf{H}^0$ is the input of the first layer, and the output node features of l^{th} layer are used as input of layer $l+1$. In this equation, each layer only considers the first-order neighbors of each node. To capture the features

of multi-order neighbors, multiple layers are needed. The final node features (also called node embeddings) can be obtained from the output of the last layer.

In the context of source identification, the GCNSI [25] method exploited the Graph Convolutional Network approach to enhance the accuracy of the LPSI method. It generated multi-dimensional node embeddings instead of integer status labels, and neighbors' features were propagated via graph convolutional operations. A multi-label classification problem was formulated to predict whether each node is a source node based on the learnt node embeddings. GCNSI also assumes the equal influence of neighbors, which will cause inaccurate results.

IV. METHOD

In this section, we propose the Motif-based Graph Convolutional Networks for Source Identification (MGCNSI) framework with consideration of various social influence of neighboring nodes. The Graph Convolutional Network based approach is adopted to identify source nodes. The node embeddings are obtained via propagation of neighbors' features, but here the difference in the social influence of neighbors needs to be considered. Although the specific propagation model is unknown, different forms of social influence are captured with structural properties in the social network. In particular, the most common higher-order structure, network motifs, are adopted to capture structural properties which are further used to extract critical neighbors of different types.

The overall architecture of MGCNSI is shown in Fig. 1, which has four parts: (a) Input layer uses the infection state of the network to generate initial node features. (b) Motif matching layer constructs a set of motif networks to capture different forms of social influence. For each node, different motif networks extract different sets of critical neighbors. (c) Motif-based graph attention networks module has two components. The motif-based GCN layers use graph convolutional networks to generate node embeddings for each motif network, and only critical neighbors of original neighbors are considered for aggregation. And the attention layer aggregates the node embeddings of different motif networks. It assigns higher weights to the informative motifs corresponding to the underlying propagation model through attention mechanisms. (d) Output layer adopts the fully-connected layer to obtain the predicted value. The details of the above four parts are introduced as follows.

A. MOTIF MATCHING LAYER

To effectively capture the patterns of common structural properties, we introduce the concepts of network motifs. Network motifs are over-represented patterns of subgraphs occurring in complex networks [39]. The studies have shown that similar motifs can be found in networks that perform similar functions [39], [40]. In social networks, similar patterns such as friends in small close groups, or different groups can also be depicted by certain motifs based on their specific functions in local network structures [43], [44]. Usually, two-node, three-node and four-node motifs are used, as shown in Fig. 2.

The neighbors could exhibit different forms of influence depending on their connectivity in the social network. Thus, the motifs can be used in this paper to select critical neighbors of different types, such as friends in the same social group. In particular, inspired by Lee et al. [76], a motif-based matrix formulation is introduced to construct a set of motif networks. Specifically, given C different motifs $\mathcal{M} = \{M_1, M_2, \ldots, M_C\}$, different motif-based adjacency matrices $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C\}$ representing the set of motif networks can be constructed. For a certain motif M_c, the motif-based adjacency matrix $\mathbf{A}_c$ is defined as follows:

$$\begin{aligned} \mathbf{A}_c &= \mathbf{A}'_c + \mathbf{P}_c \\ (\mathbf{A}'_c)_{i,j} &= \begin{cases} 0 & i = j, \\ k^c_{ij} & i \neq j. \end{cases} \end{aligned} \tag{5}$$

where k^c_{ij} is the number of times node i and j are in the same subgraph instance of motif M_c, $\mathbf{P}_c$ is a diagonal matrix, and $(\mathbf{P}_c)_{i,i} = \max_{1 \leq j \leq N}(\mathbf{A}'_c)_{i,j}$. Instead of using the identity matrix in (4), using $\mathbf{P}_c$ can make a node and its most critical neighbor equally important.

The motif can help find the interesting subgraph of the original network, i.e., the motif network. Fig. 3 illustrates the process of constructing a motif network for the triangle motif. During the process, the blue nodes and solid lines in the initial network (left) that are matched with the motif (middle) are retained in the motif network (right), and other nodes and edges are discarded. As shown in the figure, this process can find out closely connected small groups. The grey nodes are disconnected in the motif network, which means that the nodes will lose some less important neighbors. In the context of source identification, it means that on the motif network of the triangle motif, interactions in closely connected small groups are retained so that those critical neighbors in small groups are selected to identify sources. Also, different types of motif networks can select critical neighbors of different forms.

B. MOTIF-BASED GRAPH ATTENTION NETWORKS

Through the motif matching layer in the previous section, a set of C different motif-based adjacency matrices $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C\}$ has been generated. The set of motif networks represents different forms of social influence. And for each node, different motif networks extract different sets of critical neighbors. For each motif network, the Graph Convolutional Network based approach is adopted to identify source node. The node embeddings are obtained via propagation of neighbors' features, and here only critical neighbors of original neighbors are considered. The layer-wise propagation rule is as follows:

$$\mathbf{X}^{l+1}_c = \sigma(\mathbf{D}^{-1/2}_c \mathbf{A}_c \mathbf{D}^{-1/2}_c \mathbf{X}^l_c \mathbf{W}^l_c), \tag{6}$$

where σ is the activation function, and ReLU is used in this paper, $\mathbf{A}_c \in \mathbb{R}^{N \times N}$ is the motif-based adjacency

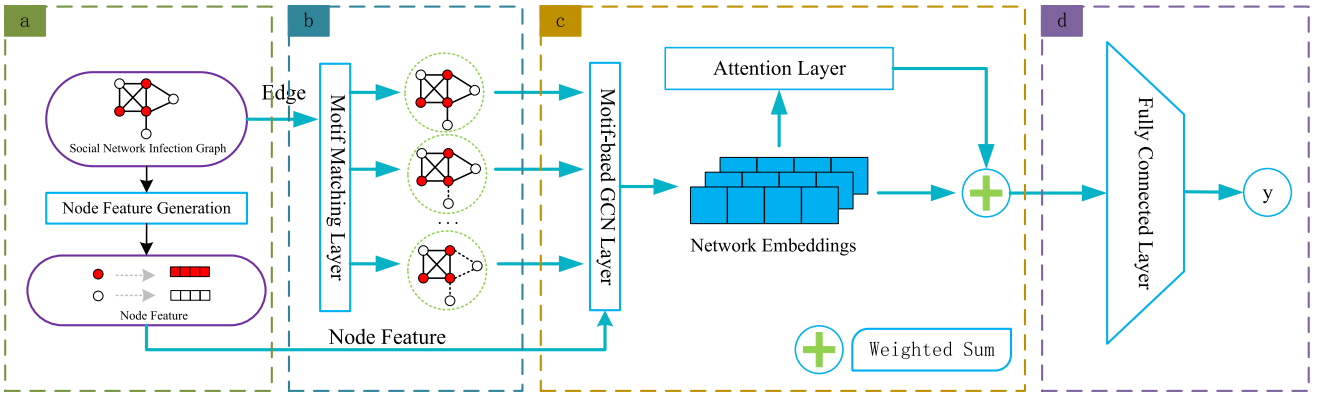


FIGURE 1. Four parts of MGCNSI: (a) Input Layer, the infection state of the network is used to generate initial node features. (b) Motif Matching Layer, a set of motif networks are constructed to capture different forms of social influence. For each node, different motif networks extract different sets of critical neighbors. (c) Motif-based Graph Attention Networks Module has two components. The Motif-based GCN layer uses graph convolutional networks to generate node embeddings for each motif network in which only critical neighbors of original neighbors are considered, and the Attention Layer aggregates them through attention mechanisms. The attention layer assigns higher weights to the informative motifs corresponding to the underlying propagation model. (d) Output Layer, fully-connected layer is adopted to obtain the predicted value.

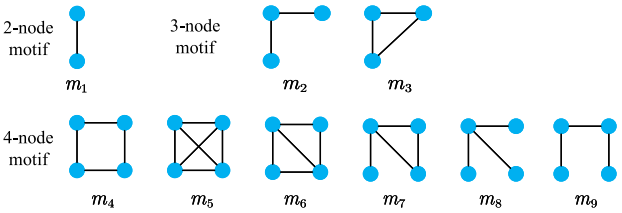


FIGURE 2. Summary of network motifs with 2-4 nodes.

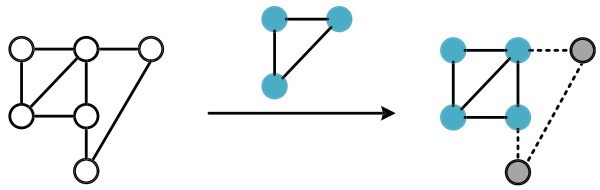


FIGURE 3. The process of constructing a motif network for the triangle motif, where the blue nodes and solid lines matched with the motif are retained, the gray nodes and dotted lines are discarded.

matrix calculated from (5), $(\mathbf{D}_c)_{ii} = \sum_j (\mathbf{A}_c)_{ij}$ represents the Laplacian matrix for $\mathbf{A}_c$, $\mathbf{X}_c^l \in \mathbb{R}^{N\times H}$ represents the node embeddings of the l^{th} layer, $\mathbf{W}_c^l \in \mathbb{R}^{H\times H}$ represents the trainable weight matrix of l^{th} layer, and H is the dimension of the hidden layer. $\mathbf{X}_c^0$ is the node features of the input layer, and the calculation method will be introduced in Section IV-C. The node embeddings obtained by GCN are the node embeddings of the last layer L, i.e., $\mathbf{X}_c^L$.

The node embeddings learnt from different motif networks are then aggregated, but the importance of them is apparently different. As the underlying propagation mechanism is unknown, to adapt to various propagation mechanisms, the attention mechanisms are adopted to automatically assign higher weights to the informative motifs corresponding to the underlying propagation model. Specifically, each motif network $\mathbf{A}_i$ has a corresponding vector $\boldsymbol{\alpha}_i \in \mathbb{R}^N$ where the attention coefficient α_{ij} represents the importance of the motif M_i for a specific node j. The attention coefficients are calculated as follows:

$$\boldsymbol{\alpha}_i = softmax(\mathbf{e}_i) = \frac{\exp(\mathbf{e}_i)}{\sum_{c=1}^{C} \exp(\mathbf{e}_c)},$$
$$\mathbf{e}_i = \sigma(\mathbf{X}_i^L \mathbf{W}_e + \mathbf{W}_b)\mathbf{W}_t \tag{7}$$

where σ represents a nonlinear function, and *tanh* is used in this paper, $\mathbf{W}_e \in \mathbb{R}^{H\times H}$ and $\mathbf{W}_b \in \mathbb{R}^{N\times H}$ are trainable weight matrices of the linear layer, $\mathbf{W}_t \in \mathbb{R}^H$ is the trainable vector which converts the matrix into a vector, and $\mathbf{e}_i$ is normalized by row-wise softmax activation to obtain $\boldsymbol{\alpha}_i$.

Eventually, the final node embeddings $\mathbf{X}_f$ can be obtained by calculating the weighted average of the node embeddings for each motif network:

$$\mathbf{X}_f = \sum_{i=1}^{C} \boldsymbol{W}_\alpha^i \odot \mathbf{X}_i^L, \tag{8}$$

where $\mathbf{X}_i^L$ represents the node embeddings corresponding to $\mathbf{A}_i$, calculated by (6), and $\odot$ represents Hadamard product. As the dimension of the node embeddings is H, we denote $\boldsymbol{W}_\alpha^i \in \mathbb{R}^{N\times H}$ as the weight matrix through concatenating the same weight vector $\boldsymbol{\alpha}_i$ H times for easier representation of (8), and $\boldsymbol{\alpha}_i$ represents the attention coefficients for $\boldsymbol{A}_i$.

C. INPUT, OUTPUT, LOSS FUNCTIONS

1) INPUTS

The original input is infection state of the network $\mathcal{Y}$. Inspired by Dong et al. [25], the features of nodes are expanded. The expanded node features are denoted as $\mathbf{Y}' = [\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4]$. The four dimensions in the i^{th} row can be treated as the four-dimensional features of node i. In particular, $\mathbf{d}_1 = \mathcal{Y}$ represents the infection state of the network, the values of infected nodes are set as 1 and those of uninfected nodes

are set as -1. $\mathbf{d}_2$ is calculated by the LPSI algorithm, that is, the output of (3), which provides initialization of features to be used by more advanced models. In order to prevent positive and negative labels from canceling each other out, $\mathbf{d}_3$ and $\mathbf{d}_4$ only consider infected nodes and uninfected nodes respectively. That is, the values for the uninfected nodes in the infection state $\mathcal{Y}$ are set to zero to calculate $\mathbf{d}_3$, and the values for the infected nodes in the infection state $\mathcal{Y}$ are set to zero to calculate $\mathbf{d}_4$. Note that the value of μ in (3) is set as 0.5, which is consistent with the original paper.

2) OUTPUTS

In the output layer, a fully-connected layer is adopted to obtain the predicted value:

$$\mathbf{y}' = \sigma(\mathbf{X}_f \mathbf{W}_f + \mathbf{b}_f), \tag{9}$$

where $\mathbf{y}'$ is the final output of MGCNSI, σ is the sigmoid function, $\mathbf{X}_f$ is the node embeddings of $\mathcal{G}$ calculated by (8), and $\mathbf{W}_f \in \mathbb{R}^H$ and $\mathbf{b}_f \in \mathbb{R}^N$ are trainable parameters of the fully-connected layer.

3) LOSS FUNCTION

The multi-source identification problem here is variance of the multi-label classification problem and it is needed to predict whether each node is a source or not simultaneously. For each given training sample, the infection state of the network in the propagation process is the input of the model, and a vector $\mathbf{y}$ denotes the ground-truth set of sources, where a value of 1 indicates that the node is the source node and 0 otherwise. And $\mathbf{y}'$ is the predicted set of sources, i.e., the output of (9). A cross-entropy loss is adopted as loss function, and L2 regularization is used to avoid overfitting. The loss function is given as:

$$\begin{aligned} \mathrm{L}(\mathbf{y}, \mathbf{y}') &= \frac{\sum_{b=1}^{B} \sum_{i=1}^{N} \mathrm{cel}(y_i^b, y'^b_i)}{B} + \lambda ||\mathbf{\Theta}||_2, \\ \mathrm{cel}(x, y) &= -x \log y - (1-x) \log(1-y), \end{aligned} \tag{10}$$

where B is the batch size, $\mathbf{\Theta}$ is the set of all the parameters of MGCNSI, $||\mathbf{\Theta}||_2$ is the L2 regularization term, and λ is the weight coefficient for the L2 regularization, with a value of 0.0001.

D. COMPLEXITY ANALYSIS

Implementing the MGCNSI framework involves two main steps: (a) constructing a set of motif networks for each network in advance; (b) main part of the framework, including generating node features, motif-based graph convolution network operations and the output module. The details are given as follows:

First of all, a set of motif networks was constructed for each network in advance using the Parallel Parameterized Graphlet Decomposition method [77]. According to [77], the computational complexity for this step is $\mathcal{O}(C\Delta^2)$, where C is the number of motifs, Δ is the maximum degree of the graph.

The main part of the framework includes four components, given as:

- Firstly, node features are generated given the original infection states of each node using the LPSI method. According to [24], the complexity for each data sample is $\mathcal{O}(N^3)$.
- For graph convolution operations in (6), according to [64], the complexity of the filtering operation is $\mathcal{O}(|E|H^2)$, and the complexity of the activation function is $\mathcal{O}(NH)$. As there are C motif networks, the overall complexity for the graph convolution operations is $\mathcal{O}(nC(|E|H^2 + NH))$, where n is the number of GCN layers.
- The attention calculation in (7) is a matrix operation with a complexity of $\mathcal{O}(CNH^2)$, and the aggregation operation in (8) is $\mathcal{O}(CNH)$. Thus the overall complexity of attention module is $\mathcal{O}(CNH^2)$.
- The output module is a fully-connected layer with a complexity of $\mathcal{O}(NH)$.

Thus, the overall complexity for each data sample in the main part is $\mathcal{O}(N^3 + nC(|E|H^2 + NH) + CNH^2)$. The run-time for datasets of different sizes is calculated. On average, MGCNSI spent 0.01 seconds, 0.011 seconds, 0.013 seconds, and 0.011 seconds for each data sample for the Dolphin, Jazz, Ego-facebook, and three synthetic networks respectively. In comparison, GCNSI spent 0.008 seconds, 0.009 seconds, 0.012 seconds, and 0.01 seconds, respectively. Thus, our method MGCNSI can improve the accuracy of the model at the expense of a reasonable increase in run-time.

V. EXPERIMENTS

In this section, we first conducted a series of experiments on both synthetic and real-world networks to evaluate the performance of our method. The results showed that the motif-based graph convolutional network framework is more accurate in identifying propagation sources. Then, we studied the influence of hyper-parameters on the experimental results and analyzed the role of different motifs in source identification.

A. DATASETS AND BASELINES

The following real-world network datasets are used in this paper:

- **Dolphin** [78] is an undirected social network of frequent associations between dolphins in a community living off Doubtful Sound, New Zealand.
- **Jazz** [79] is an undirected network of Jazz musicians.
- **ego-Facebook** [80] was collected from survey participants who used the Facebook app.

The details of the datasets are shown in Table 2.

The following methods are our model and baselines of this paper for multi-source identification:

- **MGCNSI**: Motif-based Graph Convolutional Networks based source identification model, which is the proposed model in this paper.
- **GCNSI** [25]: Graph Convolutional Networks based source identification model. This is the state-of-the-

TABLE 2. Basic information about the networks. $|V|$ is the number of nodes, $|E|$ is the number edges, $\langle k \rangle$ is the average degree, and $\langle c \rangle$ is the average clustering coefficient.

Dataset	$\|V\|$	$\|E\|$	$\langle k \rangle$	$\langle c \rangle$
Dolphin	62	159	5.13	0.26
Jazz	198	2742	27.70	0.62
ego-Facebook	4039	88234	43.69	0.61
core-periphery	872	2000	4.59	0.019
random	1000	2000	4.0	0.003
hierarchical	996	2000	4.02	0.025

art GCN-based approach to identify multiple sources when the underlying propagation model is unknown. The social influence of neighbors is assumed to be identical.

- **LPSI** [24]: Multi-source identification model based on label propagation algorithm, which is the first work to identify multiple sources without any information of the specific propagation model.
- **NetSleuth** [11]: Multi-source identification model based on minimum description length approach under the SI model.
- **Zang** [12]: Multi-source identification method based on a divide-and-conquer approach under the SIR model.

B. EXPERIMENT SETTINGS

1) GENERATION OF PROPAGATION TRACES

In previous studies, propagation traces were typically generated by IC model, SI model, and SIR model. In particular, the infection probabilities p of IC model, SI model and SIR model were sampled from the uniform distribution $U(0, 1)$, and the recovery probabilities of the SIR model were sampled from the uniform distribution $U(0, p)$. In this paper, to simulate the information propagation in social networks. In addition to the above basic propagation mechanisms, common situations in social networks mentioned in Section III-B were also considered. In particular, the infection probability between two nodes within the same community is high, whereas that for two nodes between different communities is low. And the infection probability for the two nodes within denser parts of a community is higher. Hence, instead of sampling the infection probabilities from the uniform distribution $U(0, 1)$, here the probabilities are proportional to the number of common neighbors connecting the two nodes. Besides, a node can infect its neighbor not only through their direct connection, but also through group interactions. Here the probabilities of group interactions were sampled from the uniform distribution $U(0, 1)$. We simulated a propagation process until at least 30% of the network nodes were infected.

The data was generated from different underlying propagation models with different numbers of sources. Specifically, for the basic propagation mechanisms used, the ratio of the IC model, SI model and SIR model in the datasets is 1:1:1. The ratio of common situations of information diffusion in social networks in the datasets is 1:1. Also, for datasets of different sizes, the number of source nodes is different. For datasets with a number of nodes less than or equal to 500, the number of source nodes K was set as 2, 3, and 5, while for datasets with a number of nodes greater than 500, the K was set as 3, 5, and 10. For each propagation model, the proportion of training data generated for different source numbers is also 1:1:1.

All the data were mixed and shuffled before training, thus the source identification methods could not know the specific propagation models. For the test data, propagation traces of different propagation models and different source numbers were distinguished through labels to acquire different evaluation results. The ratio of training set, validation set, and test set is 8:1:1.

TABLE 3. Optimal settings of hyper-parameters.

Dataset	learning rate	GCN layers	hidden size	dropout rate	batch size	epoch
Dolphin	0.011	2	512	0.1	1000	200
Jazz	0.011	2	512	0.1	1000	200
Ego-facebook	0.013	3	512	0.2	1500	300
core-periphery	0.012	2	512	0.1	1000	200
random	0.012	2	512	0.1	1000	200
hierarchical	0.012	2	512	0.1	1000	200

2) IMPLEMENTATION DETAILS

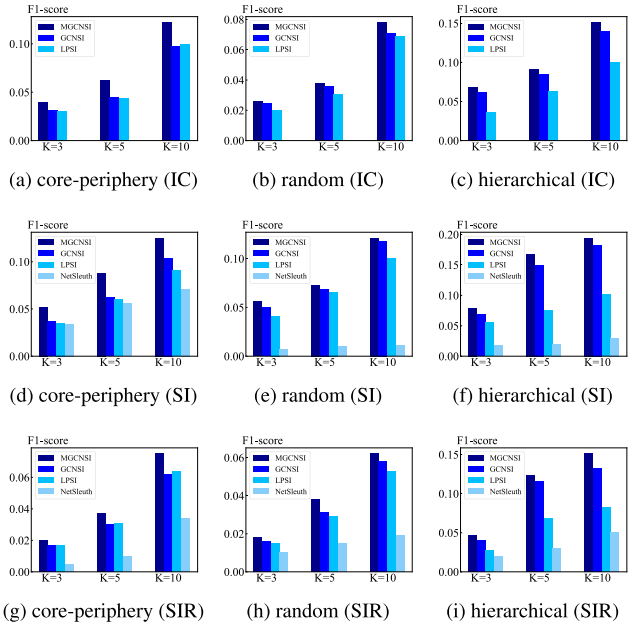
All the experiments were conducted on a Linux server with two 2.4GHz 10-core CPUs, 128GB RAM, and four NVIDIA GeForce RTX 3090 graphics cards. Pytorch[1] was used to implement our method and the Parallel Parameterized Graphlet Decomposition (PGD) library[2] was utilized to generate the motif networks in Section IV-A. The back-propagation algorithm was used to optimize the model, and the optimizer for gradient ascent is Adam [81].

The optimal values of the hyper-parameters of our model are different for different datasets, and the values of hyper-parameters were set using cross-validation. Some key validation processes of the important hyper-parameters are shown in Section V-E. Optimal settings of hyper-parameters are shown in Table 3. The parameters of the baselines are consistent with those in the original papers. The initial inputs were generated with the LPSI method, and the value of μ in (3) was set as 0.5, which is consistent with the original paper. The weight coefficient λ in the L2 regularization of loss function to prevent overfitting was set as 0.0001.

C. EVALUATION METRICS

- **F-score** is the most common evaluation criterion for multi-labeled classification tasks, as it combines precision and recall. In this paper, the β value of F-score is set to 1, that is, recall and precision have the same

[1] https://pytorch.org/
[2] http://nesreenahmed.com/graphlets/

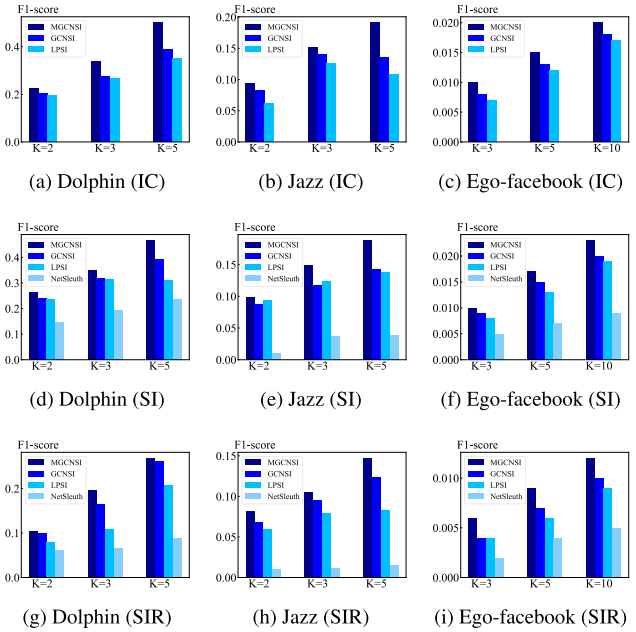


(a) core-periphery (IC) (b) random (IC) (c) hierarchical (IC)

(d) core-periphery (SI) (e) random (SI) (f) hierarchical (SI)

(g) core-periphery (SIR) (h) random (SIR) (i) hierarchical (SIR)

FIGURE 4. The results of F-score for all the methods with different numbers of source nodes K on synthetic networks. Three typical types of synthetic networks with different kinds of propagation models were considered. Compared with GCNSI, MGCNSI achieves overall improvement of about 15% on the core-periphery network and about 10% on the random and hierarchical networks. The advantage of MGCNSI is more obvious in the core-periphery network, which indicated that MGCNSI is more effective for networks with denser node neighborhoods.



(a) Dolphin (IC) (b) Jazz (IC) (c) Ego-facebook (IC)

(d) Dolphin (SI) (e) Jazz (SI) (f) Ego-facebook (SI)

(g) Dolphin (SIR) (h) Jazz (SIR) (i) Ego-facebook (SIR)

FIGURE 5. The results of F-score for all the methods with different numbers of source nodes K on real-world networks. Different kinds of propagation models were considered. Compared with GCNSI, MGCNSI achieves overall improvement of about 10%, 15%, and 13% on the Dolphin, Jazz, and Ego-facebook networks respectively. The advantage of MGCNSI is more obvious in the Jazz and Ego-facebook networks with denser node neighborhoods.

weight (F1-score). It was proved that maximizing F1-score is equivalent to maximizing the objective function in (1) [25].

- **Error Distance** is also adopted by some existing work for source prediction [53]. As is stated in [24], the results with small value of error distance are not necessarily true sources, but the nodes nearby the true sources. Nevertheless, an apparent large value of error distance can indicate that the performance is not good. To adequately evaluate the effectiveness of our model, we still show the results of error distance under the SI model. The definition of error distance is as follows:

$$d = \frac{1}{|S_g|}\Big(\sum_{i\in S_g} \min_{j\in S_p} dist(i,j) + \eta\big||S_g| - |S_p|\big|\Big), \quad (11)$$

where $|S_g|$ is the number of source nodes, $|S_p|$ is the number of predicted nodes, $dist(i,j)$ is the shortest distance from node i to node j, and for each predicted node i, the nearest source node j is taken to calculate the shortest distance. The weight η is set to 0.5.

D. RESULTS

We evaluated the performance for source identification with both synthetic and real network data sets, and compared our method with other benchmark methods. The overall performance of different methods is shown in Figs. 4 and 5.

1) EXPERIMENTS ON SYNTHETIC NETWORKS

Firstly, the performance of our method MGCNSI under different types of synthetic networks was analyzed. In particular, three typical types of synthetic networks were generated using the Kronecker graph approach [82] under the SNAP platform [83]: core-periphery networks (parameter matrix: [0.9 0.5; 0.5 0.3]) [84] which simulate the real-world networks, random networks (parameter matrix: [0.5 0.5; 0.5 0.5]) [85] used in the studies of physics, and hierarchical networks (parameter matrix: [0.9 0.1; 0.1 0.9]) [86]. The basic information of the networks can be found in Table 2.

As shown in Fig. 4, MGCNSI achieves the best performance under each kind of propagation model no matter how the number of source nodes K is set. Compared with GCNSI, the overall improvement is about 15% on the core-periphery network and about 10% on the random network and hierarchical network. The results verified the effectiveness of our method, and indicated that capturing different forms of social influence with structural properties of social networks can help better identify probable source nodes, even though the specific propagation model is unknown.

In particular, the following observations were obtained:

- The advantage of MGCNSI over other methods is more obvious in the core-periphery network, which simulates real-world networks. This is due to the structure of core-periphery networks. The edges in the random network are uniformly generated, whereas the core area of the core-periphery network is relatively denser than

the periphery area. Hence the core nodes have more neighbors and more available propagation paths, which brings difficulties to the identification of source nodes in this area. Nevertheless, our method can select important neighbors in terms of social influence by assigning higher weights to some relevant motifs automatically, and thus only critical propagation paths are considered in the motif-based graph convolutional network framework. As a result, substantial advances were observed. This indicated that MGCNSI is more effective for those networks with denser node neighborhoods.

- In addition, it can be observed that the overall source identification accuracy on the hierarchical network is higher than that on other networks, but the advantage over other methods is not obvious. This is also due to the structure of hierarchical networks. We found that the infected node sets caused by different source nodes do not overlap much, as the propagation cannot spread far from current branches. This makes the source identification problem easier for all the methods, and thus the advantage of our method is not significant.
- Besides, all the methods have poor performance in terms of the SIR propagation mechanism. We investigated the temporal traces of propagation and found a certain number of infected nodes are in the recovered state and can not be observed. As a consequence, all the methods consider these nodes as uninfected nodes, which causes inaccurate results.

2) EXPERIMENTS ON REAL-WORLD NETWORKS

Further experiments were conducted on three real-world networks of different sizes to evaluate the effectiveness of our method. The results are shown in Fig. 5. Overall, compared with GCNSI, the performance is improved by approximately 10%, 15%, and 13% in the Dolphin, Jazz, and Ego-facebook networks respectively. The results are consistent with those based on the synthetic networks, and verify that our method is more advantageous. Also, the following observations were obtained:

- More evident enhancement can be found in the Jazz and Ego-facebook networks. As shown in Table 2, the Jazz and Ego-facebook networks have higher average clustering coefficient, hence the nodes in Jazz and Ego-facebook networks have denser node neighborhoods than those in the Dolphin network. As anticipated, the MGCNSI can automatically select critical neighbors from the larger neighbor sets, and achieve evident enhancement.
- In addition, the results showed that the overall performance for all the methods is better for networks with smaller sizes. The best performance is achieved on the Dolphin network and the worst performance on the Ego-facebook network. As larger and denser networks have more available propagation paths, greater uncertainty is introduced which further makes it difficult for all the methods to identify the sources hidden in the observed infection state graph.
- Also, it is worth noting that although all the methods have relatively poor performance in terms of the SIR propagation mechanism, the improvement ratio is positively correlated with network size. Since the recovery rate remains the same, there exist more observed infected nodes in larger networks. As more observed critical neighbors can help better identify probable source nodes, the situation is somewhat alleviated for larger networks.

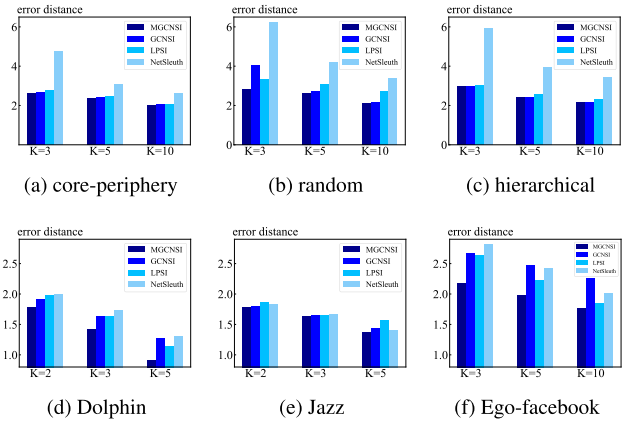
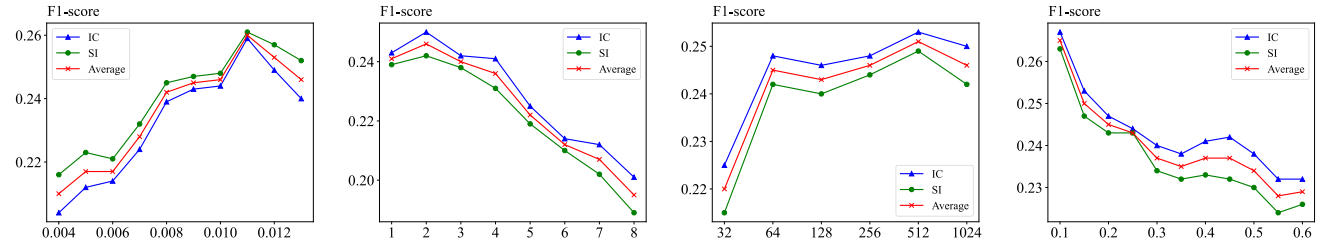


FIGURE 6. The results of error distance for all the methods with different numbers of source nodes K under the SI model. Even though the method with smaller error distance can not necessarily predict most real sources, the large error distance can indicate that the method's prediction ability is not good. In most cases, MGCNSI performs best, followed by GCNSI. The performance improvement is the most significant in the Ego-facebook network, which indicates that the advantage of our method in terms of error distance is more significant in larger networks.

3) PERFORMANCE MEASURED WITH ERROR DISTANCE

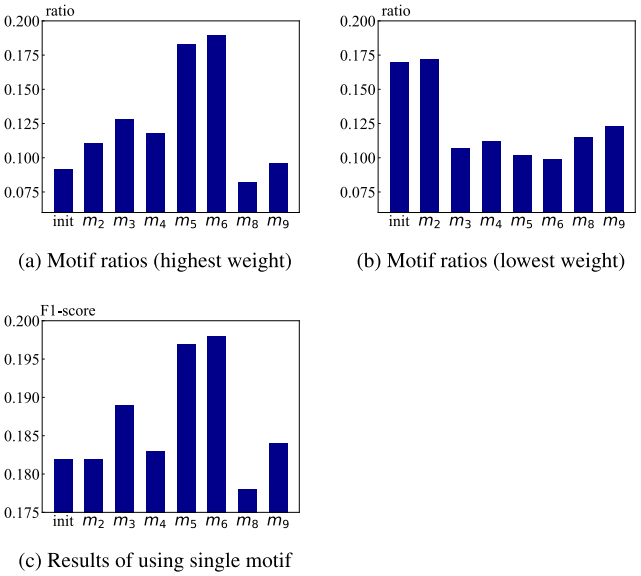
In addition, Fig. 6 shows the performance measured in terms of error distance. The error distance measures the average shortest path between predicted source nodes and real source nodes. Even though the method with a small error distance can not necessarily predict most real sources, the large error distance can indicate that the method's prediction ability is not good. Due to the fact that the value of error distance can only reflect the predictive ability to a certain extent, this paper only shows the error distance under the SI model. As shown in Fig. 6, MGCNSI achieves the best performance in all networks.

Specifically, for the synthetic networks, there is a performance improvement of about 5% on the core-periphery and hierarchical network and about 10% on the random network. In contrast with the results of F-score, the network with the most obvious advantage of MGCNSI over other methods is the random network instead of the core-periphery network. Notably, in terms of the overall error distance of all the methods, the smallest value is achieved in the core-periphery network. We would therefore recall that the F-score for the core-periphery network is not the highest. The results



(a) Learning rate (b) GCN Layers (c) Hidden size (d) Dropout rate

FIGURE 7. Impact of hyper-parameters (a) Learning rate, (b) GCN Layers, (c) Hidden size and (d) Dropout rate on the model. The optimal values of the hyper-parameters are 0.011, 2, 512 and 0.1 respectively.



(a) Motif ratios (highest weight) (b) Motif ratios (lowest weight)

(c) Results of using single motif

FIGURE 8. The ratios of nodes that put the (a) highest, (b) lowest weight on each motif network. And (c) the results of using each individual motif network separately for source identification. Among them, m_5, m_6 are very prominent, followed by m_3. The results indicated the importance of the connected groups of neighbors for source identification. Note that *init* represents the original network, and m_7 is removed since the network does not contain m_7 structure.

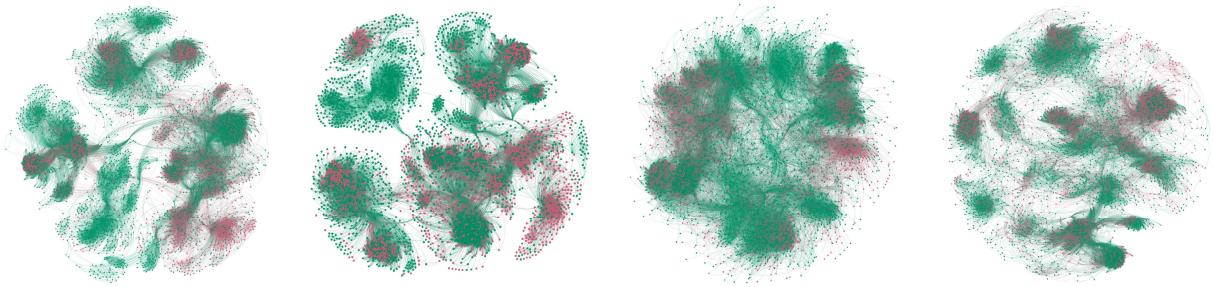
showed that the predicted source nodes for this network are not very precise, but they are near the real sources. As the core area of the network is dense, more paths among the core nodes exist. Hence, it is hard to precisely identify the sources in the core area, and the predicted source nodes are always nearby the real sources. The error distance metric prefers the methods whose predicted source nodes are close to the real sources. As a consequence, measuring the error distance is more advantageous for the core-periphery network, which simulates the real-world networks. As the error distances of baseline methods are already small, further improvement is not as significant as that in the random network.

In the real-world networks, compared with GCNSI, the performance improvement in terms of error distance in the Ego-facebook network is the most significant, which reaches 20%. It can be observed that Ego-facebook network is apparently larger than the other two datasets, and greater uncertainty is introduced for the larger networks with more available propagation paths. Also, the larger values of network diameters also result in the higher maximum value of the error distances. As a result, the overall performance of all the methods in the Ego-facebook network is obviously worse than that in the other two datasets. Meanwhile, our method can select critical neighbors so that only critical propagation paths are considered. As anticipated, the advantage of our method is most obvious in the Ego-facebook network.

E. HYPER-PARAMETER ANALYSIS

The values of hyper-parameters were set using cross-validation. Here the validation results of the important hyper-parameters including the learning rate, number of GCN layers, hidden size, and dropout rate are shown in Fig. 7. The validation on the Dolphin dataset is demonstrated as a case study, and the results on other datasets are similar. The results are given as:

- The learning rate controls the speed of convergence for the training procedure. As the value of learning rate increases, the performance of MGCNSI first increases then drops, it achieves the best performance with learning rate set as 0.011 on average. It shows that a suitable learning rate is important for the Adam optimizer.
- The number of GCN layers determines the order of neighborhoods a node can reach. The best performance is obtained with two layers, and as the number of layers continues to increase, the F-score has a tendency to decrease. Usually, too many GCN layers can lead to overfitting.
- Hidden size refers to the dimension size of the middle layer of GCN in the framework, and the results demonstrate that 512 is the best choice. If the value of hidden size is too small, the samples can not be learned well. Meanwhile, a large value of hidden size may cause overfitting.
- In order to prevent overfitting, the dropout rate is used, and the results show that the performance is best when the dropout rate is set to 0.1. The possible reason is that the Dolphin network is small and does not need a high dropout rate.



(a) m_5 (b) m_6 (c) m_8 (d) m_9

FIGURE 9. The distribution of infected nodes on different motif networks, where red nodes represent infected nodes and green nodes represent uninfected nodes. As shown in the figure, obvious clusters of infected nodes can be found for m_5 and m_6 that represent connected groups, while the motif networks corresponding to m_8 and m_9 are relatively "chaotic."

F. MOTIF ANALYSIS

1) STATISTICS

The network motifs were analyzed to better understand the role of motifs for source identification. Fig. 8a/8b shows for each motif network the proportions of nodes that put the highest/lowest weight on that motif network in the Ego-Facebook dataset. It can be observed that different motifs occupy different proportions.

As is shown in Fig. 8a, m_5, m_6 are very prominent, followed by m_3. The results demonstrate that the ''fully-connected'' motifs (m_3, m_5) representing connected groups can play an important role on the model. Motif m_6 is more prominent than m_3, possible because a m_6 motif is composed of two m_3 motifs, and two nodes on the diagonal of m_6 need to participate in two m_3 motifs at the same time, which further indicates the importance of the connected groups of neighbors.

In Fig. 8b, *init* (representing the original network) and m_2 have the highest ratios, which means they are the least important. The reason for the high ratio of nodes for m_2 is that almost all neighbors can be contained in a m_2 motif for each node, which leads to the failure of that motif to select critical neighbors.

The results of using each individual motif network separately to identify source nodes for the same dataset are shown in Fig. 8c. It can be observed that the pattern is consistent with that in Fig. 8a, that is, the results corresponding to the motifs m_5, m_6 are very prominent, followed by m_3. This further confirms our previous discussion.

2) VISUALIZATION

To further understand the role of these prominent motifs, on the Ego-Facebook dataset, motif networks corresponding to m_5, m_6, m_8, and m_9 were extracted, where m_5 and m_6 represent connected groups. Gephi[3] was used for visualization. As shown in Fig. 9, different colors are assigned to nodes to represent different node states, where infected nodes are red and uninfected nodes are green. It can be observed that in the corresponding figures of m_5 and m_6, the infected nodes are distributed in obvious clusters, which indicates that the corresponding motif network can help find ''critical'' neighbors along infection paths. Meanwhile, the motif networks corresponding to m_8 and m_9 are relatively ''chaotic'', and infected nodes and uninfected nodes are mixed together. Thus those motif networks bring difficulties to source identification.

[3]https://gephi.org/

VI. CONCLUSION

In this paper, we studied the multi-source identification problem in social networks without prior knowledge of underlying propagation model. As the specific propagation model can not be obtained, existing studies assumed the influence of each neighbor to be identical in the propagation. In reality, mesoscopic structures in social networks such as communities contain useful information of social influence, which can be utilized to better identify source nodes. Accordingly, a motif-based graph convolutional network framework was proposed. In particular, multiple network motifs were introduced to capture the patterns of common structural properties, so that different types of social influence can be depicted. Also, to adapt to various propagation mechanisms, an attention mechanism was introduced to automatically assign higher weights to the informative motifs of the underlying propagation mechanism. The major findings can be drawn as follows:

- Experiments on several synthetic and real-world networks showed that our method outperforms the state-of-the-art baselines in different situations. For the synthetic networks, it was observed that the performance improvement of our method is most significant in the core-periphery network that simulates real-world networks. As the core area of the core-periphery network is relatively denser than the periphery area, the core nodes have more neighbors. Our method is able to select critical neighbors among a large number of neighbors in terms of social influence, which helps better identify the sources. For the same reason, in real-world networks, more evident enhancement was found in networks with denser node neighborhoods.

- The role of network motifs for source identification was also analyzed. It showed that the motifs representing connected groups such as "fully-connected" motifs are more prominent, which indicated the importance of the connected groups of neighbors in the underlying propagation mechanism. It also showed that motif networks are able to help find critical neighbors along infection paths.

This paper has some limitations. There are still some interesting phenomena that need to be explored. First, all the methods have suboptimal performance under the SIR propagation mechanism. We investigated the temporal traces of propagation and found a certain number of infected nodes are in the recovered state and could not be observed, and all the methods consider these nodes as uninfected nodes. This simple assumption caused inaccurate results. Second, the propagation mechanism is assumed to be identical for different types of information. In reality, the mechanism can be different for some specific ones, such as rumors. Moreover, the network remains static in this study. Our proposed method provides a simple insight with consideration of social influence. For future work, the above assumptions could be further relaxed. Finally, some other issues such as scalability may also be further studied.

ACKNOWLEDGMENT

(Kaijun Yang and Qing Bao contributed equally to this work.)

REFERENCES

[1] J. Kostka, Y. A. Pignolet, and R. Wattenhofer, "Word of mouth: Rumor dissemination in social networks," in *International Colloquium on Structural Information and Communication Complexity*. Berlin, Germany: Springer, 2008, pp. 185–196.

[2] D. B. Kurka, A. Godoy, and F. J. Von Zuben, "Online social network analysis: A survey of research applications in computer science," 2015, *arXiv:1504.05655*.

[3] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–36, Mar. 2019.

[4] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 2671–2675.

[5] S. Shelke and V. Attar, "Origin identification of a rumor in social network," in *Cybernetics, Cognition and Machine Learning Applications*. Singapore: Springer, 2020, pp. 89–96.

[6] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?" *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5163–5181, Aug. 2011.

[7] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: Theory and experiment," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Modeling Comput. Syst.*, Jun. 2010, pp. 203–214.

[8] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample-path-based approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 408–421, Feb. 2016.

[9] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemic with a dynamic message-passing algorithm," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, no. 1, Jul. 2014, Art. no. 012801.

[10] S. J. Kazemitabar and A. A. Amini, "Approximate identification of the optimal epidemic source in complex networks," in *Proc. NetSci-X, 6th Int. Winter School Conf. Netw. Sci.*, Tokyo, Japan, 2020, pp. 107–125.

[11] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?" in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 11–20.

[12] W. Zang, P. Zhang, C. Zhou, and L. Guo, "Locating multiple sources in social networks under the SIR model: A divide-and-conquer approach," *J. Comput. Sci.*, vol. 10, pp. 278–287, Sep. 2015.

[13] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "K-center: An approach on the multi-source identification of information diffusion," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2616–2626, Dec. 2015.

[14] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "Rumor source identification in social networks with time-varying topology," *IEEE Trans. Depend. Secure Comput.*, vol. 15, no. 1, pp. 166–179, Jan. 2018.

[15] Q. Huang, C. Zhao, X. Zhang, and D. Yi, "Locating the source of spreading in temporal networks," *Phys. A, Stat. Mech. Appl.*, vol. 468, pp. 434–444, Feb. 2017.

[16] Y. Chai, Y. Wang, and L. Zhu, "Information sources estimation in time-varying networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2621–2636, 2021.

[17] R. Paluch, Ł. G. Gajewski, K. Suchecki, and J. A. Hołyst, "Source location on multilayer networks," 2020, *arXiv:2012.02023*.

[18] J. Goldenberg, B. Libai, and E. Müller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Lett.*, vol. 12, no. 3, pp. 211–223, 2001.

[19] J. E. Cohen, "Infectious diseases of humans: Dynamics and control," *JAMA, J. Amer. Med. Assoc.*, vol. 268, no. 23, p. 3381, Dec. 1992.

[20] L. J. S. Allen, "Some discrete-time Si, SIR, and SIS epidemic models," *Math. Biosci.*, vol. 124, no. 1, pp. 83–105, Nov. 1994.

[21] Y. Zhou, C. Wu, Q. Zhu, Y. Xiang, and S. W. Loke, "Rumor source detection in networks based on the SEIR model," *IEEE Access*, vol. 7, pp. 45240–45258, 2019.

[22] H. Liu, Q. Bao, H. Qiu, M. Xu, and B. Shi, "Source identification of asymptomatic spread on networks," *IEEE Access*, vol. 9, pp. 34142–34155, 2021.

[23] C. Yixin, C. Xinyue, L. Yi, W. Hanzhen, L. Yongqing, and X. Yang, "Detecting rumor dissemination and sources with SIDR model," *Data Anal. Knowl. Discovery*, vol. 5, no. 1, pp. 78–89, 2021.

[24] Z. Wang, C. Wang, J. Pei, and X. Ye, "Multiple source detection without knowing the underlying propagation model," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 217–223.

[25] M. Dong, B. Zheng, N. Q. V. Hung, H. Su, and G. Li, "Multiple rumor source detection with graph convolutional networks," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 569–578.

[26] J. Wang, J. Jiang, and L. Zhao, "An invertible graph diffusion neural network for source localization," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 1058–1069.

[27] C. Ling, J. Jiang, J. Wang, and Z. Liang, "Source localization of graph diffusion via variational autoencoders for graph inverse problems," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 1010–1020.

[28] Z. Lin, Y. Zhang, Q. Gong, Y. Chen, A. Oksanen, and A. Y. Ding, "Structural hole theory in social network analysis: A review," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 3, pp. 724–739, Jun. 2022.

[29] D. Jin, Z. Yu, P. Jiao, S. Pan, D. He, J. Wu, P. S. Yu, and W. Zhang, "A survey of community detection approaches: From statistical modeling to deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1149–1170, Feb. 2023.

[30] S. Majhi, M. Perc, and D. Ghosh, "Dynamics on higher-order networks: A review," *J. Roy. Soc. Interface*, vol. 19, no. 188, Mar. 2022, Art. no. 20220043.

[31] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[32] N. Barbieri, F. Bonchi, and G. Manco, "Influence-based network-oblivious community detection," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 955–960.

[33] I. Iacopini, G. Petri, A. Barrat, and V. Latora, "Simplicial models of social contagion," *Nature Commun.*, vol. 10, no. 1, pp. 1–9, Jun. 2019.

[34] M. Granovetter, "The strength of weak ties," *Amer. J. Sociol.*, vol. 78, no. 6, pp. 1360–1380, 1973.

[35] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 16, pp. 5962–5966, Apr. 2012.

[36] Q. Bao, W. K. Cheung, Y. Zhang, and J. Liu, "A component-based diffusion model with structural diversity for social networks," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1078–1089, Apr. 2017.

[37] S. Chowdhary, A. Kumar, G. Cencetti, I. Iacopini, and F. Battiston, "Simplicial contagion in temporal higher-order networks," *J. Phys., Complex.*, vol. 2, no. 3, Sep. 2021, Art. no. 035019.

[38] J. Fan, Q. Yin, C. Xia, and M. Perc, "Epidemics on multilayer simplicial complexes," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 478, no. 2261, May 2022, Art. no. 20220059.

[39] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.

[40] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, Feb. 2017, pp. 601–610.

[41] J. Jiang, Y. Hu, X. Li, W. Ouyang, Z. Wang, F. Fu, and B. Cui, "Analyzing online transaction networks with network motifs," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 3098–3106.

[42] X. Zhang, L. Xu, and Z. Xu, "Influence maximization based on network motifs in mobile social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2353–2363, Jul. 2022.

[43] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 555–564.

[44] A. Friggeri, G. Chelius, and E. Fleury, "Triangles to capture social cohesion," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 258–265.

[45] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Phys.*, vol. 6, no. 11, pp. 888–893, Nov. 2010.

[46] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Phys. A, Statist. Mech. Appl.*, vol. 391, pp. 1777–1787, Feb. 2012.

[47] S. Gao, J. Ma, Z. Chen, G. Wang, and C. Xing, "Ranking the spreading ability of nodes in complex networks based on local structure," *Phys. A, Stat. Mech. Appl.*, vol. 403, pp. 130–147, Jun. 2014.

[48] K. Berahmand, N. Samadi, and S. M. Sheikholeslami, "Effect of rich-club on diffusion in complex networks," *Int. J. Modern Phys. B*, vol. 32, no. 12, May 2018, Art. no. 1850142.

[49] K. Berahmand, A. Bouyer, and N. Samadi, "A new centrality measure based on the negative and positive effects of clustering coefficient for identifying influential spreaders in complex networks," *Chaos, Solitons Fractals*, vol. 110, pp. 41–54, May 2018.

[50] P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Phys. Rev. Lett.*, vol. 109, no. 6, Aug. 2012, Art. no. 068702.

[51] F. Yang, S. Yang, Y. Peng, Y. Yao, Z. Wang, H. Li, J. Liu, R. Zhang, and C. Li, "Locating the propagation source in complex networks with a direction-induced search based Gaussian estimator," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105674.

[52] C. Shi, Q. Zhang, and T. Chu, "Locating the source of diffusion in networks under mixed observation condition," *Phys. Lett. A*, vol. 434, May 2022, Art. no. 128033.

[53] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "Identifying propagation sources in networks: State-of-the-art and comparative studies," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 465–481, 1st Quart., 2017.

[54] Y. Jiang, R. Wang, J. Sun, Y. Wang, H. You, and Y. Zhang, "Rumor localization, detection and prediction in social network," *IEEE Trans. Computat. Social Syst.*, early access, Nov. 10, 2022, doi: 10.1109/TCSS.2022.3216923.

[55] L. Sun, Y. Rao, X. Zhang, Y. Lan, and S. Yu, "MS-HGAT: Memory-enhanced sequential hypergraph attention network for information diffusion prediction," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, 2022, pp. 4156–4164.

[56] N. Barbieri, F. Bonchi, and G. Manco, "Cascade-based community detection," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, Feb. 2013, pp. 33–42.

[57] Q. Bao, W. K. Cheung, B. Shi, H. Qiu, and L. Ma, "Joint learning of embedding-based parent components and information diffusion for social networks," *IEEE Access*, vol. 8, pp. 50709–50720, 2020.

[58] B. Jhun, M. Jo, and B. Kahng, "Simplicial SIS model in scale-free uniform hypergraph," *J. Stat. Mech., Theory Exp.*, vol. 2019, no. 12, Dec. 2019, Art. no. 123207.

[59] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[60] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *Proc. 2nd Int. Conf. Learn. Represent.*, Banff, AB, Canada, 2013, pp. 1–14.

[61] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 129–150, Mar. 2011.

[62] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3844–3852.

[63] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, "CayleyNets: Graph convolutional neural networks with complex rational spectral filters," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 97–109, Jan. 2019.

[64] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent.*, Toulon, France, 2017, pp. 1–14.

[65] J. Atwood and D. Towsley, "Diffusion-convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[66] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn.*, New York, NY, USA, 2016, pp. 2014–2023.

[67] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 1263–1272.

[68] X. Liu, M. Yan, L. Deng, G. Li, X. Ye, D. Fan, S. Pan, and Y. Xie, "Survey on graph neural network acceleration: An algorithmic perspective," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 5521–5529.

[69] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann, "Pitfalls of graph neural network evaluation," in *Proc. Relational Represent. Learn. Workshop, (NeurIPS)*, Montreal, QC, Canada, 2018, pp. 1–11.

[70] F. Errica, M. Podda, D. Bacciu, and A. Micheli, "A fair comparison of graph neural networks for graph classification," in *Proc. 8th Int. Conf. Learn. Represent.*, Addis Ababa, Ethiopia, 2020, pp. 1–14.

[71] O. Shchur and S. Günnemann, "Overlapping community detection with graph neural networks," in *Proc. Deep Learn. Graphs Workshop, KDD*, Anchorage, AK, USA, 2019, pp. 1–7.

[72] J. Qiu, J. Tang, H. Ma, Y. Dong, K. Wang, and J. Tang, "DeepInf: Social influence prediction with deep learning," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2018, pp. 2110–2119.

[73] F. Zhou, X. Xu, G. Trajcevski, and K. Zhang, "A survey of information Cascade analysis: Models, predictions, and recent advances," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–36, Mar. 2022.

[74] C. Hongsong, "Networks, crowds, and markets: Reasoning about a highly connected world (Easley, D. and Kleinberg, J.; 2010) [book review]," *IEEE Technol. Soc. Mag.*, vol. 32, no. 3, pp. 10–30, Fall. 2013.

[75] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 137–146.

[76] J. B. Lee, R. A. Rossi, X. Kong, S. Kim, E. Koh, and A. Rao, "Graph convolutional networks with motif-based attention," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 499–508.

[77] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield, "Efficient graphlet counting for large networks," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 1–10.

[78] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropolog. Res.*, vol. 33, no. 4, pp. 452–473, Dec. 1977.

[79] P. M. Gleiser and L. Danon, "Community structure in Jazz," *Adv. Complex Syst.*, vol. 6, no. 4, pp. 565–573, Dec. 2003.

[80] J. Leskovec and J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 539–547.

[81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–15.

[82] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, pp. 985–1042, Feb. 2010.

[83] J. Leskovec and R. Sosič, "SNAP: A general-purpose network analysis and graph-mining library," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 1, pp. 1–20, Jan. 2017.
[84] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Trans. Knowl. Discovery From Data*, vol. 5, no. 4, pp. 1–37, Feb. 2012.
[85] A. Chris, "Networks, crowds, and markets: Reasoning about a highly connected world," *Math. Comput. Educ.*, vol. 47, no. 1, p. 79, 2013.
[86] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008.

KAIJUN YANG received the B.Eng. degree in network engineering from Shaoyang University, Shaoyang, China, in 2020. He is currently pursuing the master's degree with the School of Cyberspace, Hangzhou Dianzi University, Hangzhou, China. His research interests include social network analysis, machine learning, and graph neural networks.

QING BAO received the B.Sc. degree from the Department of Computer Science and Technology, East China Normal University, Shanghai, China, in 2011, and the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2016. She is currently an Associate Professor with the School of Cyberspace, Hangzhou Dianzi University, China. Before that, she was a Post-doctoral Research Fellow with Hong Kong Baptist University. Her research interests include graph data mining, social network analysis, and health informatics. She was a recipient of the Best Student Paper Award in the 2013 IEEE/WIC/ACM International Conference on Web Intelligence. She is a reviewer of various journals and a program committee member of several conferences.

HONGJUN QIU received the B.Sc. degree in computer science and technology from Beijing Forestry University, Beijing, China, in 2003, and the Ph.D. degree in computer application from the Beijing University of Technology, Beijing, in 2010. She is currently a Lecturer with the School of Cyberspace, Hangzhou Dianzi University, China. Her research interests include complex systems/networks, autonomy-oriented computing, and health informatics.