Federation University

FedUni ResearchOnline

https://researchonline.federation.edu.au

Copyright Notice

This is the published version of:

Yu, S., Xu, J., Zhang, C., Xia, F., Almakhadmeh, Z., & Tolba, A. (2019). Motifs in Big Networks: Methods and Applications. IEEE Access, 7, 183322–183338.

Available online at https://doi.org/10.1109/ACCESS.2019.2960044

Copyright ©The Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (http://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 17, 2019, accepted November 28, 2019, date of publication December 16, 2019,
date of current version December 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960044

Motifs in Big Networks: Methods and Applications

SHUO YU[1], JIN XU[1], CHEN ZHANG[1], FENG XIA[1,2], (Senior Member, IEEE), ZAFER ALMAKHADMEH[3], AND AMR TOLBA[3,4]

[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116024, China
[2]School of Science, Engineering and Information Technology, Federation University Australia, Ballarat, VIC 3350, Australia
[3]Computer Science Department, Community College, King Saud University, Riyadh 11451, Saudi Arabia
[4]Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Shibin Al Kawm 32651, Egypt

Corresponding author: Feng Xia (f.xia@ieee.org)

This work was supported by the Deanship of Scientific Research at King Saud University under Grant RG-1438-027.

ABSTRACT Motifs have been recognized as basic network blocks and are found to be quite powerful in modeling certain patterns. Generally speaking, local characteristics of big networks could be reflected in network motifs. Over the years, motifs have attracted a lot of attention from researchers. However, most current literature reviews on motifs generally focus on the field of biological science. In contrast, here we try to present a comprehensive survey on motifs in the context of big networks. We introduce the definition of motifs and other related concepts. Big networks with motif-based structures are analyzed. Specifically, we respectively analyze four kinds of networks, including biological networks, social networks, academic networks, and infrastructure networks. We then examine methods for motif discovery, motif counting, and motif clustering. The applications of motifs in different areas have also been reviewed. Finally, some challenges and open issues in this direction are discussed.

INDEX TERMS Network motif, motif counting, motif discovery, motif clustering, network science.

I. INTRODUCTION

In molecular biology science, motifs are defined as frequent subpatterns that appear more often in a set of data, and are considered to have specific biological significance [1], [2]. If this pattern is detected in a set of DNA or protein sequences, it is a sequence motif. Structural motifs are super-secondary structures with common biological functions. Unlike molecular biology, system biology focuses more on molecular interactions than on individual molecules [3], [4]. Therefore, it shows the interaction of molecules by establishing biological networks, and the concept of network motifs comes into being. However, this concept is not limited to biological networks but also used in social networks, electronic line networks, wireless networks among several other networks.

Network science is a typical interdisciplinary subject, which focuses on the qualitative and quantitative laws of complex network systems. The scope of network science research is extensive. One of the most typical applications is to mine information in large-scale networks. For example, we can mine similarity information of scholars in academic networks, group or political tendency information of users in social networks, etc. Network science is to be distinguished from graph theory. Although it is a discipline based on graph theory, in network science, real information will occupy the main body of research. In a sense, any network can be abstracted into a graph in the field of mathematics. Conversely, any graph can be applied to a network in a real-world situation. In this paper, the two concepts (graphs and networks) are used interchangeably.

As the scale of the network increases, more recurrent and statistically significant correlation patterns appear in real networks. This kind of pattern is regarded as a **motif**. The concept of network motifs is theorized by Milo *et al.* [5] for the first time. And the network motif is a high-frequency low-order subgraph. Such subgraphs appear much more frequently in real networks than in random networks. There are two key points in the motif concept. One is the occurrence number of subgraphs, which involves the isomorphism of subgraphs. The subgraph counting problem has been recognized as a non-deterministic polynomial (NP) problem. The other one is that motifs appear more frequently in real networks. In some specific network environments, the number

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

of some motifs in a real network may be hundreds of times larger than that of a corresponding random network.

Motifs have been regarded as the basic structural unit in big networks. It has been shown that a large amount of information has not yet been excavated in the higher-order structure of big networks. Meanwhile, motifs also play significant roles in network evolution and optimization. Therefore, the mining of motifs in the network is helpful to make a deeper understanding of the network pattern. This topic has drawn researchers' attentions and developed significantly in recent years. Some studies focus on the relationships between motifs and network topology indices. Meanwhile, motif discovery, motif counting, motif clustering, as well as other related issues have increasingly drawn scholars' attentions [6], [7]. At the same time, the research of motifs is integrated with a large number of other fields, which makes the efficiency of some tasks more accurate. However, the information on motifs itself is less enriched and expanded. In other words, the concept of motifs is more inclined to a set of equivalent subgraphs.

The expansion of the motif concept can offer a more comprehensive and rich perspective for information mining in the network. Therefore, in recent years, the concept of motifs has been further expanded, which incentivizes researchers to pay more attention to the pattern characteristics of the network neutron map. This weakens the significance of its statistics. Based on the broad definition of motifs, some scholars have conducted a deep study on the interaction between agents in communication networks. Kovanen *et al.* [8] take the subject attributes (such as gender, age, etc.) into the structure of the motifs. There exists abundant studies regarding motifs, including survey papers of motifs in biological science, network science, etc. However, the current theoretical system lacks a systematic literature review of motifs. Therefore, we summarize the current studies and introduce the methods and applications of motifs. FIGURE 1 shows the structure of this paper.

In this paper, we first present the definition of motifs and some related concepts in Section II. Then we introduce networks with motif-based structures in Section III. Methods for motif relevant processing are reviewed in Section IV, including motif discovery, motif counting, and motif clustering. The applications of motifs in big networks are illustrated in Section V. We discuss challenges and open issues in Section VI. Section VII concludes the paper.

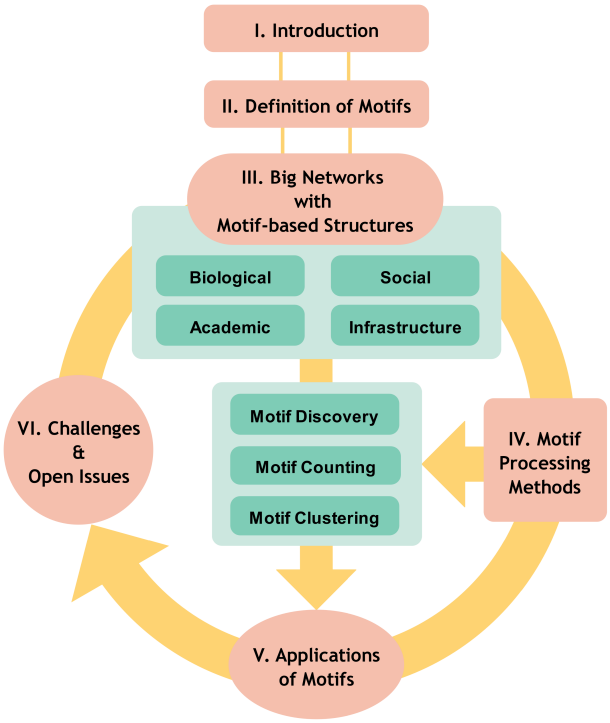


FIGURE 1. The structure of this paper.

II. DEFINITION OF MOTIF

In this section, we will discuss the definition of motifs as per the traditional field of biology and the network motif that is used in this paper. In addition, some related concepts, including subgraph isomorphism, induced subgraphs, and non-induced subgraphs, are discussed in detail in this section.

A. NETWORK MOTIF

Milo *et al.* [5] first proposed the concept of network motifs in the field of networks. A network motif is a special subgraph structure in a network. The formal definition of network motif depends on three constraints, i.e., P, U, and D.

- P is **the probability threshold** of a subgraph structure, which is used to make sure that the occurrence frequency in the real network is much higher than that in the random network;
- U is **the unique threshold**, which ensures the occurrence frequency of a subgraph structure in the real network;
- D is **the minimum difference threshold** which is used to make sure that the difference between a subgraph's occurrence frequency in the real network and that in the random network is too small.

Based on the above three constraints, the definition of the network motif is as follows:

Definition 1 (Network Motif): In a graph $G = (V, E)$ abstracted by a network, a network motif refers to an induced subgraph G_k bound by a set of parameters $\{P, U, D, N\}$. P, U and D are the probability thresholds, the unique N is the number of random networks satisfying the following three conditions:

$$Prob(\bar{f}_{rand}(G_k) > f_{real}(G_k)) \leq P,$$
$$f_{real}(G_k) \leq U,$$
$$f_{real}(G_k) - \bar{f}_{rand}(G_k) > D \times \bar{f}_{rand}(G_k).$$

wherein, $f_{real}(G_k)$ is the occurrence frequency of the subgraph in the real network, $\bar{f}_{rand}(G_k)$ is the average of the subgraph's occurrence frequency in a set of random networks. $Prob(\bar{f}_{rand}(G_k) > f_{real}(G_k))$ indicates the probability that the

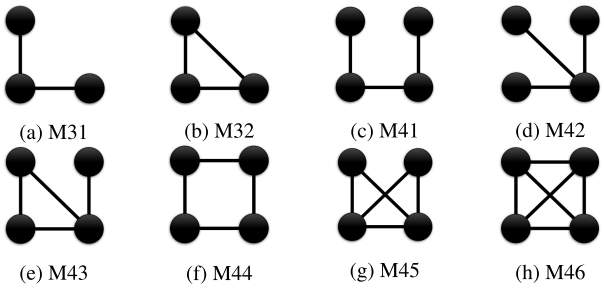


FIGURE 2. The low-order homogeneous network motif used in this paper.

occurrence frequency of the subgraph in random networks is higher than that in real networks.

The network motif is often divided into the low-order motif and the high-order motif according to the number of nodes inside it. In general, we often denote a network motif consisting of three or four nodes as a **low-order motif**, and a network motif with five or more nodes is called a **high-order motif**. Generally, methods involving higher order motifs are more time-consuming. The reason behind is that higher order motifs have more complicated structures than lower order ones. For example, for a directed network, there are 13 kinds of three-order motif structures, 199 kinds of four-order motif structures, and 9,366 kinds of five-order motif structures. It can be observed that the number of motif types from four-order to five-order has risen dramatically. And this makes it difficult for researchers to count and analyze all kinds of high-order motifs.

B. THE CHARACTERISTICS OF NETWORK MOTIFS

Typically, network motifs have the following three characteristics:

(1) A network motif generally refers to a low-order subgraph and is usually composed of 3-8 nodes. In FIGURE 2, we enumerate several typical network motifs consisting of three and four nodes. For example, the four-order fully-connected motif can be denoted as **M46**. Wherein, **4** indicates the order of the motif, that is, the number of nodes, and **6** indicates a certain structure of four-order motifs.

(2) The occurrence frequency of different network motifs can be totally different based on the network properties and motif structures.

(3) A network motif generally corresponds to a specific pattern in the real network. For example, a three-order motif in the social network corresponds to a ternary closure in social relationships..

The generality of the network motif in the real network and its corresponding practical significance will be introduced detailedly in Section III.

C. THE RELATED CONCEPTS OF NETWORK MOTIF

There are some common related concepts when dealing with motifs in the network. In this subsection, we introduce these related concepts briefly.

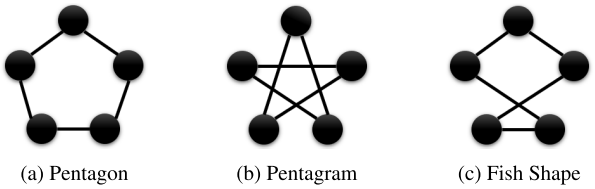


FIGURE 3. The figure of isomorphic subgraph.

1) SUBGRAPH ISOMORPHISM

Since a network motif is a kind of high-frequency low-order subgraph in the network, the subgraph isomorphism problem is often encountered during detection and counting. Scholars have studied the subgraph isomorphism problem for a long period [9]–[11]. This problem refers to the node mapping relationship between two subgraphs wherein all edges are also eligible. Specifically, two graphs can be transformed into two identical graphs by changing the relative positions of the nodes. For example, the three subfigures shown in FIGURE 3 are mutually isomorphic.

2) INDUCED SUBGRAPHS

The concepts of induced subgraphs and non-induced subgraphs are common, especially in the field of motif counting. An induced subgraph is also known as a derived subgraph, which refers to a subgraph that is separated from the original graph directly. An induced subgraph maintains all of the edge connections from the original graph. Nodes in a non-induced subgraph do not need to maintain these connections. For example, for a graph G consisting of n_G nodes, we extract a subgraph S containing n_S ($n_S \leq n_G$) nodes in G. If there are m edges among the n_S nodes in G, and all these m edges are retained in S, then S is an induced subgraph. Otherwise, it is a non-induced subgraph.

3) MOTIF COMBINATION

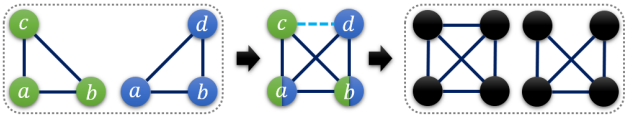
Motif combination utilize a combination of lower-order motifs to represent higher-order motifs. It is used in many algorithms to improve overall efficiency. Specifically, the motif combination method converts a high-order motif with high counting and analyzing complexity into a combination of low-order motifs. By analyzing these low-order motifs, the relevant information of the corresponding high-order motif can be obtained. As shown in FIGURE 4, (a) indicates that two three-order motifs **M32** + **M32** can be used to represent the relevant information about the four-order motif **M46** or **M45**, and (b) shows that two three-order motifs **M31** + **M32** can be used to represent information about the four-order motif **M45** or **M43**. For counting, the number of **M46**, **M45**, and **M43** can be counted based on the number of any three-order motif. Motif combination is usually used for issues involving higher-order motifs.

III. BIG NETWORKS WITH MOTIF-BASED STRUCTURES

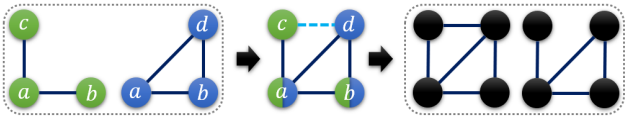
Network motifs are widely found in various types of real networks. In this section, we discuss the universality of motifs

TABLE 1. The specific data of low-order motifs in two biological networks and corresponding random networks.

Network Name	$\|V\|$	$\|E\|$	$\max(d)$	$\bar{d}$	M32	M31	M46	M45	M44	M43	M42
HS-CX[R]	4.4K	108.8K	-	-	20011.9	5205195.3	30.2	16350.8	2894135.3	722418.4	85224102.3
HS-CX			473	49	1344008	9295131	18528261	48106932	281776829	12036149	338509641
grid-yeast[R]	6K	313.9K	-	-	23799.6	8126790.6	23.5	16082.0	3664359.9	915075.9	139027182.7
grid-yeast			5.1K	104	935809	24869968	3934713	40011118	522345146	37887701	4900692083



(a) The combination of **M32 + M32** can be used to represent **M46** or **M45**



(b) The combination of **M31 + M32** can be used to represent **M45** or **M43**

FIGURE 4. Illustration of a three-order motif combination representing four-order motif.

in various real networks. Meanwhile, we demonstrate the practical meaning of different kinds of motifs. On the one hand, high-order motif calculations are complicated, and counting in large networks often consumes a lot of time. For higher order motif counting of the social networks and infrastructure networks used in this article, we can hardly predict the completion time. On the other hand, there are so many types of high-order motifs that it is difficult to distinguish which of them are of practical meaning. Some high-order motifs may represent a certain amino acid complex structure in a biological network, but may have no practical significance in an infrastructure network. Also, there exists too many heterogeneous structures of higher-order motifs, which make it difficult and time-consuming to detect them all. Therefore, we mainly discuss the high-frequency appearance characteristics and practical significance of low-order motifs, i.e., three-order and four-order motifs.

A. BIOLOGICAL NETWORKS

Generally, biological networks are small in scale but relatively complex. A biological network may be a simple undirected unweighted homogeneous network, such as some amino acid networks and protein networks. But it can also be very complex, such as gene-related networks and protein compound networks.

- From the perspective of nodes, there are various types of nodes. The most common nodes of biological networks are proteins, but still may be genes, RNA or DNA.
- From the perspective of edges, an edge of biological networks generally corresponds to the interaction between two nodes, including interactions in physics, biochemistry, or biological functions.

In the current research, scholars mainly focus on typical biological networks such as protein interaction networks [12], [13], transcriptional regulatory networks [14], [15], and metabolic networks [16].

We selected two typical bioprotein networks, *bio-grid-yeast* (http://networkrepository.com/bio-grid-yeast.php) and *bio-HS-CX* (http://networkrepository.com/bio-HS-CX.php) [17]. In addition, we also generate corresponding random networks for comparison. These random networks have the same number of nodes and edges as the corresponding real networks. For generalization, we created 10 corresponding random networks for each real network. And the final motif count results are the average values of the 10 random networks. The network information and the number of various low-order motifs are shown in TABLE 1. Wherein, $|V|$ indicates the number of nodes and $|E|$ indicates the number of edges in the network. The network labelled by "[R]" refers to the corresponding random network. Since the random networks' counting results are taken from the average of 10 networks, we did not count the maximum and average degree information (i.e., $\max(d)$ and $\bar{d}$, respectively). Results are shown in FIGURE 5 (a), (b). It can be seen that the motifs in the real biological network appear more times than that in the random network, which proves the conjecture mentioned above. It is worth noting that **M31** and **M41** are paths with three nodes and four nodes. They are only considered in very special circumstances, and are therefore not counted in the experiments.

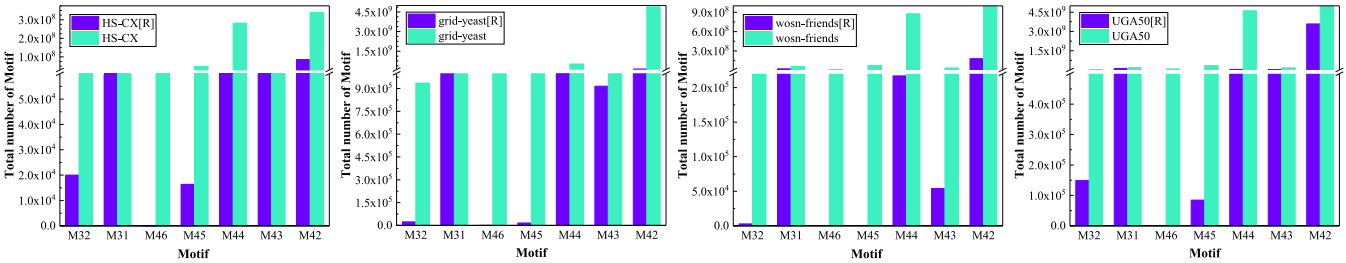
The actual meaning of the motif we find in the biological network is consistent with the original definition of a motif in the biological domain. For example, recognized motifs in protein networks correspond to a specific combination of proteins that make up a practical role in forming a protein complex.

B. SOCIAL NETWORKS

There are many family maps and interpersonal diagrams in real life, which can be regarded as the prototypes of social networks. In recent years, social communication has become common. Modern communication tools such as e-mails, blogs, and instant messaging platforms have brought new relationship patterns to social networks [18], [19].

In general, social networks have larger scales than biological networks. Meanwhile, social networks contain a wealth of short paths. Generally, social networks have the following properties.

- From the perspective of node types, the node type of social networks is relatively simple. Nodes are always used to represent social users, entities, institutions, or groups, etc;



(a) BIO-HS-CX (b) BIO-grid-yeast (c) SO-wosn-friends (d) SO-UGA50

FIGURE 5. The number of low-order motifs in two biological networks and two social networks.

TABLE 2. The specific data of low-order motifs in two social networks and corresponding random networks.

Network Name	$\|V\|$	$\|E\|$	$\max(d)$	$\bar{d}$	M32	M31	M46	M45	M44	M43	M42
wosn-friends[R]	63.7K	1.3M	-	-	2820.4	20945578.9	0.0	44.0	216944.6	54007.0	178922156.6
wosn-friends			1.8K	39	3501542	60606726	13333952	75983448	876646309	35788148	3420344309
UGA50[R]	24.4K	1.2M	-	-	148642.6	112587229.0	56.0	84368.3	42591047.2	10657402.3	3584603746.9
UGA50			2.9K	96	10087811	179395190	73219381	340961294	4620177217	158812569	16878134535

- From the perspective of edges, the connection between two nodes can represent the social relationship between users, institutions, etc.Edges in the social networks can be weighted or unweighted.

Typical social networks include email networks, Facebook social networks, Google social networks, group relationship networks, etc.

To explore motifs in social networks, we use two Facebook social networks in our experiments, which are named *socfb-UGA50* (http://networkrepository.com/socfb-UGA50.php) and *socfb-wosn-friends* (http://networkrepository.com/socfb-wosn-friends.php) [17]. Similar to the experimental process of the above biological network, we generate 20 corresponding random networks and count the low-order motifs appearing in these networks. The experimental results are shown in FIGURE 5 (c), (d), and TABLE 2. We can also see that motifs are ubiquitous in social networks. In random networks, the four-node fully-connected motif **M46** rarely (or even does not) appears, but there exists a large number of such motifs in real networks.

C. ACADEMIC NETWORKS

Academic networks and social networks have many similarities in some network features. Academic networks can be classified into various kinds, including cooperation networks, citation networks, etc. Related studies primarily focus on applying academic networks to solve certain practical problems, including identifying scientific research teams, recommending academic entities, analyzing current research hotspots, predicting knowledge transfer trends, evaluating scholars' research directions, etc., [20]–[22].

In general, the structures of academic networks are more complicated than social networks. Similarly, analyzing academic networks, the following properties can be observed.

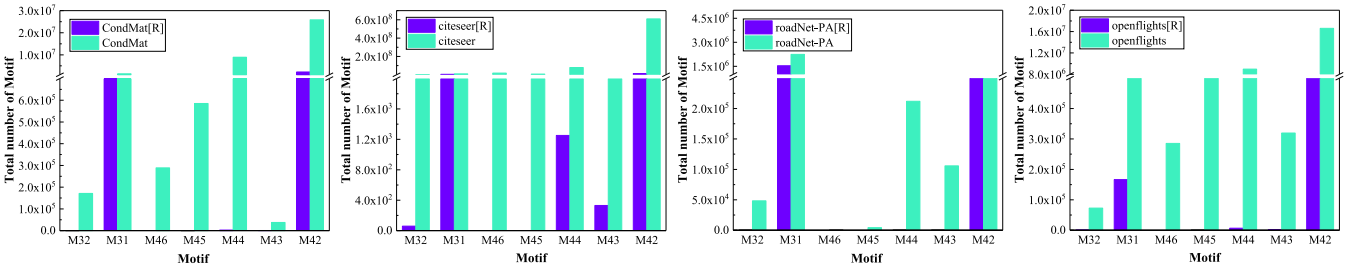
- Nodes in academic networks are mainly researchers, publications, venues, research institutions, etc. Each type can be further divided. For example, researchers can be further divided according to professors and students, and papers can be divided based on subject, keywords, disciplines, etc.
- Relationships in academic networks are more complicated and diverse. The edges in a network can be used to represent the relationships between papers and researchers, the citing relationships between one paper and another, the subordination relationships of researchers and institutions, the co-authorships among researchers, inter alia.

Consequently, most of the academic networks under investigation are heterogeneous networks.

We use two academic collaboration networks to verify the existences of motifs in academic networks. The names of the two real networks are *ca-citeseer* (http://networkrepository.com/ca-citeseer.php) and *ca-CondMat* (http://networkrepository.com/ca-CondMat.php) [17]. The experimental results are shown in FIGURE 6 (e), (f), and TABLE 3. The two academic networks examined in this paper are relatively sparse. It is obvious that **M46**, i.e., four-order fully connected motif and **M45**, i.e., four-order chord ring motif rarely exists in random networks. However, in real networks, such motifs widely exist. This illustrates that motifs are indeed universal in academic networks.

D. INFRASTRUCTURE NETWORKS

Infrastructure networks are a special kind of networks. Normally each infrastructure network is transformed from a set of physical infrastructures. Many real networks can be considered as typical infrastructure networks. The aircraft shipping network employs airports as nodes and aircraft routes as



(e) CA-CondMat (f) CA-citeseer (g) IN-roadNet-PA (h) IN-openflights

FIGURE 6. The number of low-order motifs in two collaboration networks and two infrastructure network.

TABLE 3. The specific data of low-order motifs in these two collaboration networks and corresponding random networks.

Network Name	$\|V\|$	$\|E\|$	$\max(d)$	$\bar{d}$	M32	M31	M46	M45	M44	M43	M42
CondMat[R]	23.1K	93.4K	-	-	106.8	779751.0	0.0	0.5	2740.4	664.8	2219535.1
CondMat			279	8	171051	1446763	289216	585398	8897769	37757	25868047
citeseer[R]	227.3K	814.1K	-	-	58.7	5830727.5	0.0	0.0	1252.5	331.2	13919162.3
citeseer			1.4K	7	2713298	9729730	18716073	7658658	79269580	82822	609279988

TABLE 4. The specific data of low-order motifs in two infrastructure networks and corresponding random networks.

Network Name	$\|V\|$	$\|E\|$	$\max(d)$	$\bar{d}$	M32	M31	M46	M45	M44	M43	M42
roadNet-PA[R]	1.1M	1.5M	-	-	0.2	1534367.5	0.0	0.0	0.7	0.3	721410.0
roadNet-PA			9	2	48319	2246191	14	4195	211896	105769	984583
openflights[R]	2.9K	30.5K	-	-	203.7	166424.7	0.0	12.8	6455.5	1615.4	585774.1
openflights			473	20	72852	639476	285560	1466065	8970572	319408	16602431

edges. The urban transportation network takes road junctions in cities as nodes and connected roads as edges. The power transmission network represents nodes as stations, power plants or power centers and edges as cable carrying power. Hence studying infrastructure networks can obtain meaningful information so that the administrators can take immediate response to emergency events, such as high-risk areas of aviation, traffic jams, and areas with dangerous power loads [23].

The network structure of infrastructure networks is related to the actual infrastructure. Although it is generally a sparse network, its network size varies greatly. For example, the node number of an aircraft route network is generally less than 10,000. But for a transportation network, in contrast, it can have millions or even tens of millions of nodes.

Based on the above observation, we select two different infrastructure networks for experiments, namely *inf-openflights* (http://networkrepository.com/inf-openflights.php) [17] for the air transport route and *inf-roadNet-PA* (http://networkrepository.com/inf-roadNet-PA.php) [17] for real traffic. The experimental procedure remains the same. The results obtained are shown in FIGURE 6 (g), (h), and TABLE 4. Since the original real traffic network *inf-roadNet-PA* is too sparse, it is almost impossible to find four-order motifs except the four-order claw motif **M42** in the corresponding random networks obtained by randomization. The number of **M42** is affected by the three-order path **M31**, which has almost no practical significance, and it is also prominent in other random networks. For other four-order motifs, they appear more frequently in the real traffic network. This is consistent with the feature that motifs have a higher frequency in real networks than in random networks.

IV. MOTIF PROCESSING METHODS

In this section, we mainly introduce some methods for processing motifs in big networks, including three different issues, i.e. motif detection, motif counting, and motif clustering. It is worth noting that although the methods have certain similarities, their outputs and targets are quite different.

A. MOTIF DISCOVERY

Motifs are considered to be special structures with potential functions in complex networks. However, motif detection is a time-consuming task, especially in big networks. To reduce the computational complexity, scholars begin to study how to improve the efficiency of motif discovery algorithms [24], [25]. Many scholars have put their efforts into handling this task from serialized algorithms to parallel algorithms, and sampling algorithms which might lose accuracy in exchange for time [26], [27]. Under many circumstances, motifs are first to be discovered and then counted. Therefore, in the following two subsections, we introduce motif discovery algorithms and motif counting algorithms selectively based on their application scopes.

1) SERIAL ALGORITHMS

a: MotifCut

This is a motif detection method based on graph theory [28], and one of the purposes of MotifCut development is the detection of high-order motifs. MotifCut divides the input sequence into several long subcolumns, and uses these subcolumns as vertices to construct a graph. The edges between the vertices represent the similarity between them. Moreover, the densest subgraph in this graph consists of a large number of nodes that are close to each other. The square method does not assume the structure of modules and requires them to be in good agreement with the model. In dense subgraphs, those strongly dependent node pairs can also be found by MotifCut. In the first step of the algorithm, the input sequence is divided into subsegments of thousands of lengths, which contain the possible sequences. These subsegments are used as vertexes to construct a graph, and the edges of any two nodes are weighted by a function [29]. Generally, the higher the similarity between the two nodes is, the greater the edge weight value between them is. The background distribution is used to calculate the random occurrence probability of two subsegments in the input sequence. For a pair of two nodes that rarely appears in the background sequence, their edges are weighted upward. Under these conditions, the motif can be regarded as a set of nodes with high degree of deduction, which can be clearly seen by the weight of the edge in the graph. In this way, the problem can be transformed into finding the densest subgraph.

b: MotifNet

This is an open access website that analyzes motifs [30]. The user enters the network diagram with a node number or edge number, and the website can find the motifs consisting of up to eight nodes. The website can graphically output the motifs found by the query. At the same time, it can view the examples of the motifs in the network diagram. Users can also select a special node to find the subgraph centered on the node. MotifNet is based on FANMOD, so it is suitable for the highest eight-order network motif structure. This makes it possible to solve most of the motif detection problems.

c: G-trie

Ribeiro and Silva [31] proposed a G-trie data structure for storing subgraph collections. It is regarded as one of the most efficient motif discovery algorithms known at present. Using G-trie and other data structures, a set of subgraphs is stored in a prefix tree. The G-trie structure uses a common prefix of the subgraph. Each trie node corresponds to a subgraph node. The connection information between the node and its prefix node is stored in each tire node, and a path from the root node of the trie tree to the leaf node represents a single subgraph. In G-trie algorithm, a method of symmetric burst breaking is proposed, to avoid the problem of repeated subgraph enumeration caused by automorphism of subgraphs. By adding the constraint of the size of enumerated subgraph, each subgraph can only be enumerated once. The concrete step consists of three procedures. First, G-trie generates all isomorphism forms of the subgraph. Then, G-trie keeps adding constraints until there are no remaining isomerism subgraphs that meet the conditions. Finally, G-trie stores these added constraints in the nodes corresponding to the G-trie structure. By constraining the entire subgraph's search process, each subgraph can be enumerated only once. In G-trie algorithm, nauty algorithm is also used as subgraph isomorphism decision algorithm [32]. Like other algorithms, G-trie can be applied in higher order motif discovery. Similarly, the computational complexity grows as the motif order increases.

2) PARALLEL ALGORITHMS

The computational complexity of subgraph enumeration grows exponentially when the subgraph size increases. Though existing algorithms can enumerate small subgraphs quickly, the execution time can be quite long for large subgraphs. An efficient way to speed up the motif discovery algorithm is to raise the enumerating efficiency of the subgraph in parallel.

a: Parallel G-trie Algorithm

Ribeiro *et al.* [33] proposed a parallel G-trie algorithm. It has been proved that parallel G-trie can further improve the computational efficiency. In this algorithm, each node in the recurrent search tree is regarded as a unit of work, and each unit of work is independent. In order to make the algorithm adapt better to parallel needs, they adopt a distributed parallel control method. That is, the process has completed its work task and randomly requests a work task from other processes. At the same time, the authors proposed a method based on diagonal segmentation to complete the assignment of initial work tasks [34], [35]. This is to achieve the effect of load balancing.

b: ITERATIVE MAPREDUCE ALGORITHM

Vartika *et al.* [36] proposed an iterative motif discovery algorithm. This algorithm requires three independent MapReduce iterations, including ESU (Enumerate Subgraphs), calculation of subgraph labels, and aggregation of results. The first iteration is to enumerate the scale of the subgraph by ESU. The second iteration procedure is to calculate the subgraph label, which mainly aims at generating a unique label for each subgraph. Finally, the aggregation of the results mainly calculates the significance of each subgraph type and identifies the network motifs. Comparing with the modular calculation of a single node, the acceleration ratio of this method can be up to 37 times.

c: GPU-BASED PARALLEL MOTIF DISCOVERY

Lin *et al.* [37] proposed this algorithm to accelerate motif discovery algorithms. In this algorithm, subgraph matching in the process of motif discovery is paralleled by GPU, so as to save calculation time. The experimental results show that this algorithm is more effective in two magnitudes.

3) SAMPLING ALGORITHMS

With the increasing scale of subgraphs, the time required exponentially increases for enumerating subgraphs by accurate statistical analysis. To solve this problem, a sampling algorithm is proposed [38], [39]. By sacrificing the accuracy of statistics, the enumeration of subgraphs can save significant calculation time.

a: EFFICIENT SAMPLING ALGORITHM

ESA (Efficient Sampling Algorithm) was first proposed by Kashtan *et al.* [38], and is a method based on edge sampling. The main step of the algorithm is to first randomly selecting an edge from the network. This is to form an initial subgraph. Secondly, an edge is randomly selected from the edge-set that is adjacent to the subgraph. This chosen edge is then added to the constructed subgraph. Finally, the first two procedures are stopped when the subgraph has reached the specified size. However, this sampling method has been proved to be biased in the research by Wernicke [40]. That is, the chosen probabilities of each subgraph in the network are not equal. Moreover, the core subgraph is easy to be oversampled, and the same subgraph may be repeatedly drawn.

b: G-TRIE SAMPLING ALGORITHM

The G-trie sampling algorithm uses a recurrent search process [39]. The whole search process can be represented by a recurrent tree. The main idea of the G-trie sampling algorithm is to retrieve the branches of the recurrent tree with a certain probability. All the subgraphs can be found by the ESA algorithm. Compared with ESA and other algorithms, the G-trie sampling algorithm has the advantage of sampling subgraph groups.

B. MOTIF COUNTING

In the process of studying motif properties, the frequency of motifs is crucial. This is because almost all the metrics and practical indicators are based on the frequency of motifs in the network. We use the most classic indicator, *Z*-Score, and the more commonly used indicator, Abundance as examples. Both of these two indicators need to calculate the frequency of the particular motifs in the real network and in a set of corresponding random networks. To calculate the exact motif's frequency in the network, we must firstly count the number of motifs in the network. Therefore, determining the number of motifs (or the frequency of motifs) is one of the most fundamental tasks. In this subsection, we refer to these two issues collectively as the motif counting problem.

The motif counting problem refers to the counting of a particular subgraph. It is one of the important issues studied by researchers before the concept of ''network motifs'' was proposed. Similarly, the motif counting problem is also a non-deterministic polynomial (NP) problem. The greatest difficulty in solving this problem is the high computational complexity. It is well known that the traditional subgraph counting problem is to traverse all combinations of nodes in the network and try to find that if the combination constitutes a target subgraph. Wherein, the node combination that needs to be traversed has the same size as the target subgraph. If we want to count a subgraph S consisting of n_S nodes in a graph G consisting of n_G nodes, there are $C_{n_G}^{n_S}$ node combinations that need to be traversed (C represents the combinatorial number). The time cost of this algorithm is obviously too high. Although the order of the network motif is relativly low (i.e., the value of n_S is small). The counting efficiency of high-order motifs is still weak in big networks.

A variety of researchers have performed in-depth and meticulous work on motif counting. In order to cope with the bottleneck of the subgraph counting problem, researchers created a variety of pruning algorithms, prediction algorithms, motif-specific counting algorithms, sampling frequency approximation algorithms, etc. These resolution algorithms have evolved with the development of network science and network motifs. At the same time, its efficiency is improving. However, no matter what kind of solution is adopted, the core solution of the contemporary motif counting algorithm is based on whether the researcher needs the exact number of motifs in the network. The accurate methods can achieve the exact value with authenticity and uniqueness, but the time required for calculation is high. Conversely, the approximate methods run with greater speed, but the shorter the run, the less accurate the results are. Herein, we discuss these two kinds of motif counting methods in detail.

1) ACCURATE MOTIF COUNTING METHODS

When the motif concept was first proposed, the algorithm proposed by the researchers was optimized on the traditional subgraph counting algorithm for a long time. They reduced the number of subgraphs that needed to be traversed by pruning and predictive operations, thereby reducing the time complexity of the algorithm.

a: FLEXIBLE PATTERN FINDER

In order to optimize the traditional subgraph counting method, Schreiber and Schwöbbermeyer [41] proposed the FPF (Flexible Pattern Finder) algorithm. This is a tree-based pattern algorithm, in which the tree is comprised of nodes representing different patterns. The graph represented by the child nodes of the tree is obtained by adding an edge to the graph represented by the current node. In other words, the parent node's graph must represent a subgraph of the current node's graph. If only the pattern tree is used, the time complexity of the algorithm has no advantage. So in this paper, the authors define three frequency concepts and determines whether the frequency of the current pattern is below a predetermined threshold during the traversal. If it is below the threshold, the current pattern can be considered as an infrequent subgraph. The path will be abandoned, and the algorithm will not continue to traverse that part of the tree. The FPF algorithm uses this technique to complete the enumeration process, which can increase the efficiency of the

traversal algorithm. Meanwhile, since there are no restrictions on the subgraph objects it looks for, FPF can count high-order motifs. However, because the algorithm is related to the size of the search subgraph, the counting efficiency is low.

b: ENUMERATE SUBGRAPHS

Based on the mfinder algorithm, Wernicke [42] established the model of precise enumeration algorithm ESU (Enumerate Subgraphs) and frequency estimation algorithm Rand-ESU. Based on these two algorithms, a visual model tool FANMOD is established [43]. To this day, FANMOD is still one of the most famous counting tools. Also, there are still countless scholars who choose to compare their methods with FANMOD in the process of researching subgraph counting and motif counting. The exact enumeration algorithm ESU is based on enumerating the subgraph of size k, which can be implemented in a recursive process:

- At initialization, the algorithm adds a node v_a from the graph G to a node set Ω, and finds its neighbor node set $N = N_{v_a}$;
- Then, a node v_b is extracted from N and added to Ω iteratively. We update N and define it as the union of the neighbor nodes of each node in Ω;
- If the size of Ω is k, we stop the algorithm and obtain the result.

Through the above recursive process, an ESU tree can be obtained. From the ESU tree, we can retrieve all subgraph structures of size k in the network efficiently. Since the size k of the search subgraph is up to eight, the ESU algorithm can perform counting for both low-order motifs and high-order motifs. This is the exact enumeration algorithm ESU, and the frequency estimation algorithm Rand-ESU which modified it will be introduced later.

c: GROCHOW-KELLIS

Grochow and Kellis [44] proposed the Grochow-Kellis algorithm, which is also an exact algorithm for enumerating subgraphs. Its enumeration process is based on a given motif, which is denoted as a query graph. This method is a motif-centric approach that searches the whole network for all possible mappings. Wherein, the mappings are from the given motif of the big network. Although the use of mapping instead of enumeration can improve efficiency, it still costs too much time while analyzing isomorphic relationships. To reduce the cost of time, the authors created a ''symmetry breaking condition''. This condition destroys the mapping relationships of isomorphic subgraph queries, thereby eliminating redundant isomorphic decisions and greatly improving the efficiency of the algorithm. The Grochow-Kellis algorithm does not limit the size of the query graph, which means it is applicable in high-order motif counting tasks.

d: RAGE

This algorithm is proposed by Marcus and Shavitt [45], and provides a new approach for the solution of the motif counting problem. The Rage algorithm utilizes the structural features of three-order motifs and four-order motifs to compute the number of such low-order induced subgraphs and non-induced subgraphs in big networks efficiently. Rage uses the connection relationship to construct a three-order motif symbol array *NodeArray*. Based on this, the algorithm can calculate the number of non-induced subgraphs of the four-order type. Moreover, it also illustrates the relationship between the four-order motif inducing subgraph and the non-inducing subgraph. This can be used to calculate the exact number of four-order motifs efficiently, while other methods require more time. The Rage algorithm has two main disadvantages, i.e., the discrimination of isomorphic subgraphs and the counting of low-order motifs. However, its speedy operation efficiency can cover these deficiencies.

The rage algorithm has been considered to be the most effective low-order motif counting algorithm for a long time. However, the human pursuit of efficiency is endless. On one hand, some scholars have further tried to improve the efficiency of the Rage algorithm. Ahmed *et al.* [46] proposed a new perspective, that is, using the combination relationship between motifs to speed up the operation even further. This algorithm utilizes the combination relationship between four-order motifs and greatly reduces the computational complexity compared with Rage. Meanwhile, the algorithm itself can perform parallel optimization, which further improves the efficiency of motif counting. This idea has been widely adopted. On the other hand, some scholars want to calculate the number of motifs above four-order by keeping the efficiency of the Rage algorithm. The algorithm established by Pinar *et al.* [47] extended it to five-order motifs.

2) FREQUENCY ESTIMATION METHODS

At times, the most frequent motif structure is required, instead of the quantity of each motif type. Under these circumstances, scholars have found that the overall frequency of occurrence is required. Therefore, estimation methods are proposed by sacrificing partial accuracy by using sampling methods.

a: MFINDER

The Mfinder algorithm was proposed by Kashtan *et al.* [48]. It is a method based on edge random sampling, which aims at estimating the concentration of large subgraphs in a big network. Mfinder does not consider subgraph isomorphism, sampling the same subgraph, etc. However, its time complexity is unexpectedly low. The algorithm's computation time is asymptotically independent on the size of the network. The Mfinder approach significantly reduces the computation time of motif counting. This was a noteworthy achievement at the time. In summary, although Mfinder can only estimate the concentration of motifs, and mainly counts low-order motifs, it provides a more efficient way compared with the traditional exhaustive search algorithm by edge sampling.

b: RAND-ESU

As mentioned earlier, the Rand-ESU algorithm is a frequency estimation algorithm modified on the basis of the exact

TABLE 5. Typical Accurate Counting Methods and Frequency Estimation Methods.

Method Type	Name	Induced or Non-Induced	Highest Order
Accurate Counting Method	FPF	Induced Subgraph	Unlimited
	ESU	Induced Subgraph	8
	Grochow-Kellis	Both	Unlimited
	Rage	Both	4
Frequency Estimation Method	mfinder	Induced Subgraph	5
	Rand-ESU	Induced Subgraph	8
	MODA	Both	Unlimited

enumeration algorithm ESU. Although the ESU algorithm is efficient, this argument is only compared to the traditional exhaustive algorithm. In practical applications, the time cost of the ESU algorithm is still unacceptable in the face of big networks. Therefore, Wernicke [42] established the Rand-ESU algorithm based on ESU. The core idea of the Rand-ESU algorithm is to explore only a portion of the ESU tree. However, the algorithm must simultaneously ensure that it reaches every leaf node with the same probability. In order to meet this condition, the algorithm introduces a probability $p_d \in [0, 1]$, which indicates that each node of the ESU tree is expanded with the probability p_d. It can be simply proved that this method can sample all subgraphs of size k randomly and uniformly without deviation. When $p_d = 0$, the ESU tree will not expand and the algorithm will not retrieve any valid results; and when $p_d = 1$, it is equivalent to the ESU algorithm. By adjusting the probability p_d, we can adjust the sampling of the algorithm. Since the Rand-ESU method has been developed from the ESU method, it can also perform counting for both low-order motifs and high-order motifs.

c: MODA

MODA is an estimation algorithm developed on the basis of the Grochow-Kellis algorithm, which was first proposed by Omidi *et al.* [49]. MODA uses a hierarchical structure called ''extension tree'', combined with the frequency characteristics of the FPF algorithm and the motif growth method. This algorithm avoids enumerating subgraphs that are not likely to constitute a target motif while querying motifs. Based on this, a special sampling method is used to accelerate the running time of the algorithm. The advantage of MODA is that it can more effectively identify high-order motifs including those above eight nodes. At the same time, in addition to inducing subgraph information, MODA can extract non-induced subgraph information of motifs in the network. However, the speed of this algorithm still does not exceed Rand-ESU algorithm.

In addition to the motif enumeration methods and the motif frequency sampling estimation methods mentioned above (as shown in TABLE 5), scholars have combined many other techniques for the motif counting problem, resulting in various solutions. These solutions include the algorithm combined with color-coding techniques [50], MHRW (Metropolis-Hastings Random Walk) [51], MASS (Multi-Agent Spatial Simulation) [52], among others.

All the above-mentioned algorithms are proposed for counting motifs in static networks. However, there exists various network environments, such as dynamic networks (e.g., Internet), probability networks (e.g., biological probability networks), etc. Motif counting algorithms suitable for specialized networks have also been proposed [53]. Schiller *et al.* [54] proposed a flow-based algorithm that can be used to compute undirected four-order motifs in dynamic graphs efficiently. Todor *et al.* [55] also proposed an algorithm, which can be used to calculate the expected value and the variance of the key motif numbers in a biological probability network. Luo *et al.* [39] designed an effective measurement method for biological networks based on motifs, namely, LCNM (Large-scale Co-regulatory Network Motifs).

C. MOTIF CLUSTERING

Motif clustering has been regarded as high-order clustering in big networks. Clustering motifs are significant in various areas, especially in biological science. Clustering is the task of classifying objects according to their features, and the objects with similar features will be grouped together. The main process is to extract element characteristics and put the elements with similar characteristics into the same category. In general, n attributes or features of samples are taken, and are then mapped to n-dimensional vectors. In this way, the original discrete data can be converted into a graph or matrix that can be further processed by a computer. Clustering and classification operations are quite different. Classification utilizes supervised learning to divide objects into different categories according to some certain training data. Whereas, clustering simply brings similar objects together through the attributes and connections between the data, which is an unsupervised learning method.

At present, clustering generally has the following steps.

- The first step is feature extraction. According to the data features and clustering purpose, we extract the underlying features of the data. These features should be as independent as possible, and the redundant information should be minimized;
- Similarity measuring is the second step. The clustering algorithm provides the standard measurement of similarity between sample nodes through a similarity matrix or other parameters. The similarity is subsequently calculated using extracted features. This measure determines the clustering effect and clustering result;

- The third and most critical step is a clustering process. A clustering algorithm determines how the samples are classified during the clustering process;
- The final phase is result testing. The subclass results obtained by clustering algorithm are judged, and their correctness is verified;

Though the motif clustering algorithms have some common properties, different algorithms are suitable in different circumstances. Herein, we introduce motif clustering according to their application scenes, i.e., motif clustering in biological science and network science.

1) BIOLOGICAL SCIENCE

There are two significant processes of motif clustering in biological science, i.e., finding the similarities among biological sequences and detecting motifs. Obvious advantages can be gained by adopting the above two clustering methods.

a: AGGLOMERATIVE HIERARCHICAL CLUSTERING

This is a commonly applied algorithm for clustering a group of elements into subgroups or smaller clusters [56]. It categorizes the elements by joining the top two similar elements and creates a tree to represent nesting group events. This method first computes the fully aligned similarity matrix using a dynamic programming algorithm. Subsequently, the top two similar elements involved in the same motif with the highest alignment score are joined recursively. Through this joining event, the algorithm creates a new motif as an alignment of two motifs. Meanwhile, the distance matrix scores are updated according to the average distance between the motif cluster and the other motifs.

b: CyClus3D

CyClus3D is a motif clustering algorithm which is proposed for three-order motif clustering [57]. For a given three-order motif, its edges can be of any type. We assume that all motifs can be represented by a three-dimensional array T. If there exists a motif among node i, j, and k, then $T_{i,j,k} = 1$. Otherwise, $T_{i,j,k} = 0$. The motif cluster that consists of three node sets (X_1, X_2, X_3) can be defined by an aggregation score, which is shown in Equation (1).

$$S(X_1, X_2, X_3) = \frac{\sum_{i\in X_1, j\in X_2, k\in X_3} T_{i,j,k}}{|X_1|^{\frac{1}{p}}|X_2|^{\frac{1}{p}}|X_3|^{\frac{1}{p}}} \tag{1}$$

wherein, $|X_i|$ is the number of nodes in X_i. To maximize S, we need to find a real vector (x_1, x_2, x_3) that maximizes Equation (2).

$$R(x_1, x_2, x_3) = \frac{\sum_{ijk} T_{ijk} x_{1,i} x_{2,j} x_{3,k}}{\|x_1\|_p \|x_2\|_p \|x_3\|_p} \tag{2}$$

wherein, $\|x_i\|_p = (\sum_i x_i^p)^{\frac{1}{p}}$ is the p norm. The optimal value of R can be calculated by solving Euler-Lagrange equation.

In order to find a motif cluster with a high score, CyClus3D removes the non-existing motifs repeatedly until no redundant motif exists in the motif set. This algorithm can be regarded as generalization form of two-dimensional spectrum clustering algorithm. However, CyClus3D can only cluster motifs consisting of three nodes, which greatly limits its scope of application.

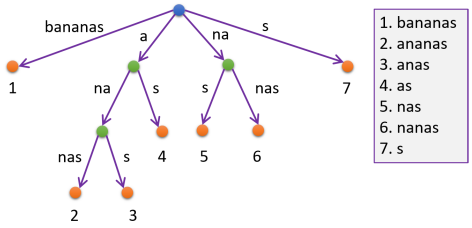


FIGURE 7. The standard suffix tree of the string "bananas".

c: SUFFIX TREE

The suffix tree can be used in the motif analysis process [58]. However it has traditionally been used in the text index method [59]. A suffix tree satisfies the following conditions. Figure 7 illustrates a standard suffix tree for the string "bananas".

(1) There is a one to one correlation from root node to leaf node for sting s. That is, each path uniquely represents a suffix of string s;
(2) Each edge represents a non-empty string;
(3) All internal nodes (except root nodes) have at least two child nodes.

All the sequences form a large class at the root node. Each time we advance one character, the sequences are divided once. Therefore, this kind of division is generally divided by the size of the character set. The path length e reflects that the length of the prefix string of the current partition is e. We assume that a horizontal dotted line is drawn at e, then the number of crossover points is the number of clusters. When the path length reaches w from this node, it is shown that the prefix sequence traversed at this time is all possible motifs. Different paths represent different subsequences, which means the number of sequences in each set is one. At this time, it is possible to continue traversing to multiple leaf nodes. One leaf node represents a suffix sequence. If there are multiple leaf nodes, these suffix sequences contain multiple identical modular instances. Theoretically, this division process can go on until some sets are empty. In fact, the goal of this clustering algorithm is to obtain a class of the right size, so this traversing can be ended quickly.

2) NETWORK SCIENCE

The purpose of local graph clustering methods is to find a cluster of nodes by exploring a small region of the graph. These methods are generally advanced in clustering around a given seed node. However, the current local graph clustering methods are not designed to explain the critical high-order structures in the network, and they cannot effectively deal with directional networks. In network science, clustering can

be divided into two classifications, i.e., node clustering and graph clustering.

a: MOTIF-BASED RANDOM WALK

This is a motif clustering method that is based on Markov process. If we visit a graph randomly, we may probably get a loop rather than go out of the graph. This loop area is a group and the route that visits the graph is always generated by a Markov process. Based on motif structures, the random walk can be guided with bias. According to the above ideas, the access to the nodes in the graph is often more likely to be concentrated in a dense area of the graph. In other words, the paths within the cluster are often abundant, but the paths between the clusters are very few. As a result, the probability of accessing other nodes in the cluster is very small. Even so, we can still access other clusters of the graph [60]. Similarly, for weighted graphs, this tendency is reflected in the weight processing of the edge. The edge weight between the nodes in the same cluster is much higher than that between the nodes in a different cluster. The edge weight should be included as the communication between weight distribution and clustering.

b: MOTIF-BASED MARKOV CLUSTERING ALGORITHM

The Markov clustering algorithm terminates access by blocking a part of the path in the Markov chain. The algorithm will adjust the transfer value so that the closer nodes can achieve higher access probability. For each node, this algorithm changes its transmission value to improve its probability to reach a closer neighborhood. This adjustment is to raise a single-row index to a non-negative high value, then set it to a normal value. It can be called ''inflated''. In a Markov chain, this process is also called ''extension''. Since the calculation of the probability depends on the number motifs that the node is included in, the probability of use is biased. There are two probability adjustment methods for the expansion: one is to enhance the current value, the other is to weaken the current value. In other words, the algorithm makes the current strong value stronger and makes the current weak value weaker. The above process is generally controlled by a parameter that ultimately affects the granularity of the group itself.

c: MOTIF-BASED APPROXIMATE PERSONALIZED PAGERANK

This is a new class of local graph clustering methods, which can solve some problems in the field of motif clustering by integrating higher-order network information [61]. The general process of this algorithm is as follows: given a graph G and a certain motif M, the algorithm aims at finding a set of nodes Ω with suitable motif conductance. Wherein, the motif conductance is extended by the traditional definition of edge conductance. It is also defined as the ratio of the cutting motifs number to the total motifs number in the graph G. The motif that is often used to calculate motif conductance is **M32**. In addition, a theory based on the node neighborhood is also developed to discover the set containing low motif conductance in this paper. Moreover, the results of these sets are used to find a good seed node to be used as the input of Motif-based Approximate Personalized PageRank algorithm.

Network motifs are widely used as structural features for biased random walks, PageRank, unsupervised learning, etc. Therefore, the order of the network motif is often not limited. Low-order motifs are more employed instead of high-order motifs in a motif cluster. This is due to the fact that in real networks, the number of high-order motifs is often scarce when compared to low-order motifs, and high-order motifs are of relatively less correspondence to specific practical meanings.

V. APPLICATIONS OF MOTIFS

With the development of motif research, the concept of motifs has been widely used in empirical networks, including international trade, Internet, transportation, inter-social virus transmission, biological systems, to name a few. These studies have proved that the application of motifs in different fields can provide an effective research perspective for the pattern characteristics of subjects in a network. Meanwhile, these applications also pay attention to the proportion of different morphological motifs in the network, and the participation of subjects in multi-morphological motifs. In this section, we will introduce the applications of motifs in different kinds of big networks, including social networks, biological networks, world trade networks, network protocols, traffic networks, Bayesian networks, Support Vector Machine and graph learning.

A. SOCIAL NETWORK ANALYSIS

When researchers study topology and complex network relationships, network motifs can reflect an intrinsic link between network nodes [62]. For some tasks, such as network-centric measurements, or time-consuming network cluster searches, combining motifs as an efficient way to analyze social networks has become a critical breakthrough. Researchers believe that using small or medium-sized networks to investigate social networks will be more effective and reliable. In fact, they have already applied network motifs to social network analysis for some time. For instance, some researchers use motifs to analyze email communication in social networks. Z-score is employed as an indicator to analyze the distribution of motifs, which is shown in Equation (3).

$$Z_m = \frac{n_M - \langle n_M^{rand} \rangle}{\sigma_M^{rand}} \tag{3}$$

wherein, n_M is the occurrence number of motif M in the real network, n_M^{rand} is the mean of M's occurrence and σ_M^{rand} is the standard deviation of M's occurrence in some corresponding random networks.

Z-score is a key parameter that reflects the importance of motifs, and it firstly requires researchers to determine the number of random networks needed to be detected. A typical application is implemented in an email-based social network. Wherein, the weight of the edge in an email-based social network is the intensity of communication. We enumerate

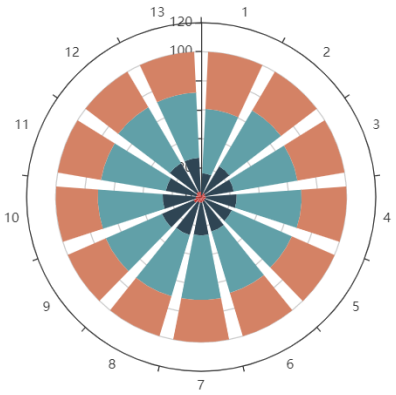
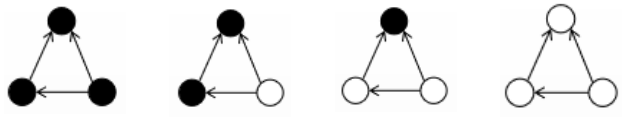


FIGURE 8. The distribution of colored network motifs in e-mail-based social network.



(a) 3B0W (b) 2B1W (c) 1B2W (d) 0B3W

FIGURE 9. Some examples of the colored motifs.

all the colored network motifs of the email-based social network, whose distributions are shown in FIGURE 8. There are 13 groups of motifs and each fan-shaped part represents a different kind.

In FIGURE 8, the outermost layer represents the distribution of the 3-black motif, the second outer layer represents that of the 1-white 2-black motif, the third layer represents that of the 2-white 1-black motif, and the innermost layer represents that of the 3-white motif. We list four examples of the four kinds of motifs mentioned above in FIGURE 9. Wherein, ''W'' indicates that several nodes in the motif are white, and ''B'' indicates that several nodes in the motif are black.

B. BIOLOGICAL APPLICATIONS

Motifs were first proposed in biological science. As such, network motifs are widely used in the biological field. Xu *et al.* [63] previously reconstructed various protein complexes in biological networks by analyzing network strucutre. Kim *et al.* [64] used motifs to find the essential protein in the network by a centrality algorithm called MCGO (Motif Centrality edGegO), which can measure the centrality of a subgraph.

Meier *et al.* [65] studied the features of brain networks, and used the entropy of directed phase transfers to search for the motif of effective connections in the network. Compared with random networks, they found the four-order motif, bidirectional two-hop path, and the three-order motif, occur frequently in the analyzed network. Moreover, they also divided the effective connectivity network into two main clusters. It was found that there exists closer connections in the posterior and the anterior region.

Sporns and Kötter [66] also studied motifs in brain networks. They found that the brain network structure is simple to figure out by network motifs. So, they compared the motif properties of real brain networks and identified the maximum number of functional motifs. Additionally, they obtained the motif frequency spectra for the matrix of C.elegans. Wherein the motif size does not exceed four. They also implemented similar experiments in other brain networks, including macaque visual cortex, macaque cortex, cat cortex, etc.

Friedman *et al.* [67] used the directed network motif to study AD (Alzheimer's Disease) and MCI (Mild Cognitive Impairment) in human networks for the first time. Wherein, the frequency of specific directed network motifs could be used to distinguish between the brain network with AD and the normal brain network. It has been proved that low-order motifs provide a useful function to distinguish the two types of patients. That is, the normal AD patients and the AD patients converted from MCI.

Motifs are applied in gene regulatory networks as well. In order to effectively select a binding sequence from a high-dimensional position, Middendorf *et al.* [68] used a modern large-margin machine learning method. Rather than over-expressed in promoter sequences, they learned functional motifs and predictive motifs. Finally, they used this model to address specific biological questions, such as restricting attention to a particular target gene.

C. WORLD TRADE NETWORK

The topological structure of World Trade Network (WTN) depends heavily on the Gross Domestic Product (GDP) of countries in the world. Thus, it is interesting to acknowledge WTN's topological properties. On one hand, researchers employed a triadic motif and analyzed its occurrence in WTN [69]. Based on the maximum-likelihood estimation of maximum-entropy models of the network, they found that reciprocity has a great influence on the motif structure, and the value is significantly high in WTN. On the other hand, some researchers also focused on analyzing the subnetworks of WTN. The evolution of WTN is also studied, including degree distribution, center coefficient, and topological structure [70], [71]. Sliding windows as a classical method can be used to find the core motif of a WTN.

Network motifs are applied in preventing a world trade crisis as well. Considering a global economic crisis, vast majority of work is focusing on the whole financial system. On the contrary, theoretical works are lightly performed regarding specifics which effectively describe the financial crisis. Saracco *et al.* [72] use motifs of bipartite networks to research the countries' product network. A part of bipartite networks represent the countries layer, and the other part represents the product layer. Wherein, various kinds of motifs are introduced, including V-motifs, Λ-motifs, X-motifs, W-motifs, and M-motifs. Saracco *et al.* find that the information of V-, Λ-motif is rather limited, but X-, W-, and M- motifs appear with a high frequency.

D. NETWORK PROTOCOLS

Some scholars use visual motifs to classify encrypted traffic. Wright *et al.* [73] used dynamic time warping of the motifs to classify encrypted traffic, and gained the heatmap for HTTP, SMTP, AIM, and SSH. On this basis, they identified the non-linear behaviors of those graphs, and the sorted graphs in order of similarity to the template connection. Finally they distinguished unauthorized servers and identified traffic generated by different processes, such as video, email, game, etc. With visual motifs in the processed heatmap, various application protocols can be identified.

E. TRANSPORTATION NETWORK

A transportation network is formed by a large number of traffic stations and lines, through the interaction of material information and energy. It plays a significant role in national economic development. Motifs are employed in analyzing the transportation network, because motifs are found to be general in this type of networks [74]. Jin *et al* [75] studied Chinese airline networks constructed by 37 airline companies. They found that motifs can reflect the structural and functional characteristics of airline networks. Moreover, the adjustment of the motif number can be applied in optimizing the overall structure of airline networks.

In an urban public transportation system network, motifs also appear frequently and can be used as a feature to study the network. Ping *et al.* [76] used the Urban Public Transportation Network (UPTN) of China to discuss the efficiency and topological nature of China's urban public transportation system. After investigating the traffic networks of the ten largest cities in China, they found that all of the topological structure of UPTN are characterized by the small-world phenomenon. Finally, they found that topological features and motifs are related to the dynamics of UPTN.

F. BAYESIAN NETWORK AND SUPPORT VECTOR MACHINE

Motifs are also applied in the classification of time series with Bayesian Network (BN) and Support Vector Machine (SVM). The task of classifying time series is crucial [77], [78]. Buza and Schmidt-Thieme [79] used motifs to improve the accuracy of classification, and they created a new motif class called generalized semi-continuous motifs. Their experimental results show that the use of motif classification models can greatly improve the performance of SVM and BN. Moreover, the machine learning algorithm realizes the fast and efficient discovery of motifs in the network [80]. For example, Kovanen *et al.* [8] combined this concept with temporal networks for mining dynamic evolution characteristics. Motifs are also well suited for network embedding tasks and achieve better performance than traditional methods [81].

G. GRAPH LEARNING

In network representation learning, network motifs can be used to depict higher-order similarities among nodes in a multivariate structure [81], [82]. This is because a network motif can capture the complex relationship of multivariate structures in real networks. Network motifs have been increasingly employed in network representation learning approaches. Pal *et al.* [83] proposed that the network motif can be applied as a structural feature to the learning process of graph convolutional networks. Likewise, Ktena *et al.* [84] presented a motif-based concept of graph similarity as well as an approach for training better motif-related variants. Latterly, network motifs have been proved to be efficient in network representation learning. Monti *et al.* [85] proposed a motif-based graph convolutional network learning method, i.e., MotifNet, which integrates the information of local network motifs within the graph convolutional network training process. MotifNet has been proved to outperform state-of-the-art graph convolutional network methods.

VI. CHALLENGES AND OPEN ISSUES

Network motifs play an important role in network science, biology science, etc. It has been proved that network motifs are the basic blocks of networks. Meanwhile, motifs provide mesoscopic thinking in big networks. Though motifs have been applied in various scenes, there still exists many challenges.

- **Computational Complexity.** The algorithms of motifs are generally time-consuming, especially when the target motifs grow larger. With the increase of vertices in motifs, related algorithms will become more complicated. As mentioned before, there are 13 kinds of directed subgraphs of 3 vertices, 199 kinds of subgraphs of 4 vertices, and 9,366 kinds of subgraphs of 5 vertices. Moreover, when these subgraphs are superimposed on a network in more potential ways, there will be more possible motifs. GPU methods have been employed to accelerate the algorithms. However, the speciality of motifs makes it difficult to implement parallel algorithms.
- **Higher-order Motifs.** Higher-order motifs are proved to have real applications in various fields. Most previous algorithms focus on small scale motifs, i.e., three-order motif and four-order motif. However, with the increasing scale of networks, higher-order motifs generally appear with a higher frequency. Therefore, greater effort should be put into developing algorithms of higher-order motif discovery, counting, profiling, and other related algorithms.
- **Heterogeneous Motifs.** In reality, real-world networks are mostly heterogeneous. Existing methods of motifs mainly focus on homogeneous motifs. This is mainly because homogeneous motifs are relatively simpler than heterogeneous motifs. Meanwhile, heterogeneous motifs are difficult to detect or cluster due to complexity and heterogeneity. Further studies may need to focus on heterogeneous motifs more.
- **Visualization Tools.** There exists abundant visualization tools for analyzing motifs such as, R package

seqLogo, MotifStack, STAMP [86], etc. However, most of the existing visualization tools are developed for biological science. Besides, there is a lack of visualization tools in other research areas, such as social relationship research, academic partner recommendation, and transportation system optimization. Therefore, motif visualization tools should be developed for more disciplines.

VII. CONCLUSION

Network motifs were widely applied in various scenes. In this paper, we focused on the relevant algorithms and applications of motifs. Firstly, we present the formal definition of motifs. Secondly, we illustrate the generality of motifs in many networks, including biological networks, social networks, academic networks, and infrastructure networks. We find that motif-based structures exist in including but not limited to these networks. Thirdly, we introduce some motif algorithms such as motif discovery, motif counting, and motif clustering. The applications of motifs in big networks have been detailed from the network perspective. Furthermore, we introduce the various applications that motifs are applied in. Finally, we discuss the challenges and open issues regarding motif applications and relevant methods. This work summarizes the current studies systematically and provides some potential directions for future work.

REFERENCES

[1] D. M. Wolf and A. P. Arkin, "Motifs, modules and games in bacteria," *Current Opinion Microbiol.*, vol. 6, no. 2, pp. 125–134, 2003.
[2] M. Altaf-Ul-Amin, F. M. Afendi, S. K. Kiboi, and S. Kanaya, "Systems biology in the context of big data and networks," *BioMed Res. Int.*, vol. 2014, pp. 1–11, May 2014.
[3] N. Pržulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: Scale-free or geometric?" *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.
[4] P. R. Somvanshi and K. Venkatesh, "A conceptual review on systems biology in health and diseases: From biological networks to modern therapeutics," *Syst. Synth. Biol.*, vol. 8, no. 1, pp. 99–116, 2014.
[5] 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, 2002.
[6] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, Mar. 2017.
[7] T. Mathew, M. Hayes, H. Fenn, H. Cutbert, C. Carter-Jones, R. Rakhit, and T. Lockie, "Impact of an acute coronary syndrome (acs) specialist nurse service to reduce time to coronary angiography+/-revascularisation," *Future Healthcare J.*, vol. 6, p. 15, Mar. 2019.
[8] L. Kovanen, M. Karsai, K. Kaski, and J. Kertész, and J. Saramäki, "Temporal motifs in time-dependent networks," *J. Stat. Mech., Theory Exp.*, vol. 2011, no. 11, 2011, Art. no. P11005.
[9] D. Eppstein, "Subgraph isomorphism in planar graphs and related problems," in *Graph Algorithms and Applications I*. Singapore: World Scientific, 2002, pp. 283–309.
[10] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1367–1372, Oct. 2004.
[11] S. Samsi, V. Gadepally, M. Hurley, M. Jones, E. Kao, S. Mohindra, P. Monticciolo, A. Reuther, S. Smith, and W. Song, "Static graph challenge: Subgraph isomorphism," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Sep. 2017, pp. 1–6.
[12] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng, "Nemofinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 106–115.
[13] N. S. Latysheva, M. E. Oates, L. Maddox, T. Flock, J. Gough, M. Buljan, R. J. Weatheritt, and M. M. Babu, "Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer," *Mol. Cell*, vol. 63, no. 4, pp. 579–592, 2016.
[14] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genet.*, vol. 31, no. 1, pp. 64–68, 2002.
[15] M. L. Arrieta-Ortiz, C. Hafemeister, A. R. Bate, T. Chu, A. Greenfield, B. Shuster, S. N. Barry, M. Gallitto, B. Liu, and T. Kacmarczyk, "An experimentally supported model of the bacillus subtilis global transcriptional regulatory network," *Mol. Syst. Biol.*, vol. 11, no. 11, pp. 1–17, 2015.
[16] V. Lacroix, C. G. Fernandes, and M.-F. Sagot, "Motif search in graphs: Application to metabolic networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 3, no. 4, pp. 360–368, Oct. 2006.
[17] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. AAAI*, 2015. [Online]. Available: http://networkrepository.com
[18] Z. Wang, Y. Zhang, Y. Li, Q. Wang, and F. Xia, "Exploiting social influence for context-aware event recommendation in event-based social networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.
[19] P. Li, H. Dau, G. Puleo, and O. Milenkovic, "Motif clustering and overlapping clustering for social network analysis," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2017, pp. 1–9.
[20] Z. Ning, L. Liu, S. Yu, and F. Xia, "Detection of four-node motif in complex networks," in *Proc. Int. Conf. Complex Netw. Appl.* Cham, Switzerland: Springer, 2017, pp. 453–462.
[21] X. Kong, M. Mao, J. Liu, B. Xu, R. Huang, and Q. Jin, "TNERec: Topic-aware network embedding for scientific collaborator recommendation," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Oct. 2018, pp. 1007–1014.
[22] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, "Academic social networks: Modeling, analysis, mining and applications," *J. Netw. Comput. Appl.*, vol. 132, pp. 86–103, Apr. 2019.
[23] X. Guan, C. Chen, and D. Work, "Tracking the evolution of infrastructure systems and mass responses using publically available data," *PLoS ONE*, vol. 11, no. 12, 2016, Art. no. e0167267.
[24] P. Ribeiro, F. Silva, and M. Kaiser, "Strategies for network motifs discovery," in *Proc. 5th IEEE Int. Conf. E-Sci.*, Dec. 2009, pp. 80–87.
[25] T. Swati and D. Soumitra, "Efficient motif discovery," *Int. J. Innov. Advancement Comput. Sci.*, vol. 7, no. 3, pp. 806–808, 2018.
[26] Z. R. M. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E. S. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, and A. Masoudi-Nejad, "Kavosh: A new algorithm for finding network motifs," *BMC Bioinf.*, vol. 10, no. 1, p. 318, Dec. 2009.
[27] S. Khakabimamaghani, I. Sharafuddin, N. Dichter, I. Koch, and A. Masoudi-Nejad, "Quatexelero: An accelerated exact network motif detection algorithm," *PLoS One*, vol. 8, no. 7, 2013, Art. no. e68073.
[28] 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*, 2017, pp. 555–564.
[29] H. A. Soufiani and E. Airoldi, "Graphlet decomposition of a weighted network," in *Proc. Artif. Intell. Statist.*, 2012, pp. 54–63.
[30] I. Y. Smoly, E. Lerman, M. Ziv-Ukelson, and E. Yeger-Lotem, "Motifnet: A Web-server for network motif analysis," *Bioinformatics*, vol. 33, no. 12, pp. 1907–1909, 2017.
[31] P. Ribeiro and F. Silva, "G-tries: An efficient data structure for discovering network motifs," in *Proc. ACM Symp. Appl. Comput.*, 2010, pp. 1559–1566.
[32] X. Li, D. S. Stones, H. Wang, H. Deng, X. Liu, and G. Wang, "Netmode: Network motif detection without nauty," *PLoS ONE*, vol. 7, no. 12, 2012, Art. no. e50093.
[33] P. Ribeiro, F. Silva, and L. Lopes, "Parallel discovery of network motifs," *J. Parallel Distrib. Comput.*, vol. 72, no. 2, pp. 144–154, 2012.
[34] T. Wang, J. W. Touchman, W. Zhang, E. B. Suh, and G. Xue, "A parallel algorithm for extracting transcriptional regulatory network motifs," in *Proc. 5th IEEE Symp. Bioinf. Bioeng. (BIBE)*, Oct. 2005, pp. 193–200.
[35] H. Wang, N. Li, J. Li, and H. Gao, "Parallel algorithms for flexible pattern matching on big graphs," *Inf. Sci.*, vol. 436, pp. 418–440, 2018.

[36] V. Verma, P. P. Kwon, and W. Kim, "Iterative Hadoop MapReduce-based subgraph enumeration in network motif analysis," in *Proc. IEEE 8th Int. Conf. Cloud Comput.*, Jun./Jul. 2015, pp. 893–900.

[37] W. Lin, X. Xiao, X. Xie, and X. Li, "Network motif discovery: A GPU approach," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 3, pp. 513–528, Mar. 2017.

[38] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, 2004.

[39] J. Luo, L. Ding, C. Liang, and N. H. Tu, "An efficient network motif discovery approach for co-regulatory networks," *IEEE Access*, vol. 6, pp. 14151–14158, 2018.

[40] S. Wernicke, "A faster algorithm for detecting network motifs," in *Proc. Int. Workshop Algorithms Bioinf.* Berlin, Germany: Springer, 2005, pp. 165–177.

[41] F. Schreiber and H. Schwöbbermeyer, "Frequency concepts and pattern detection for the analysis of motifs in networks," in *Transactions on Computational Systems Biology III*. Berlin, Germany: Springer, 2005, pp. 89–104.

[42] S. Wernicke, "Efficient detection of network motifs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 3, no. 4, pp. 347–359, Oct./Dec. 2006.

[43] S. Wernicke and F. Rasche, "FANMOD: A tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, 2006.

[44] J. A. Grochow and M. Kellis, "Network motif discovery using subgraph enumeration and symmetry-breaking," in *Proc. Annu. Int. Conf. Res. Comput. Mol. Biol.* Berlin, Germany: Springer, 2007, pp. 92–106.

[45] D. Marcus and Y. Shavitt, "Efficient counting of network motifs," in *Proc. IEEE 30th Int. Conf. Distrib. Comput. Syst. Workshops*, Jun. 2010, pp. 92–98.

[46] 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.

[47] A. Pinar, C. Seshadhri, and V. Vishal, "ESCAPE: Efficiently counting all 5-vertex subgraphs," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1431–1440.

[48] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. (2002). *Mfinder Tool Guide*. [Online]. Available: http://www.weizmann.ac.il/mcb/UriAlon/download/network-motif-software

[49] S. Omidi, F. Schreiber, and A. Masoudi-Nejad, "MODA: An efficient algorithm for network motif discovery in biological networks," *Genes Genet. Syst.*, vol. 84, no. 5, pp. 385–395, 2009.

[50] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. C. Sahinalp, "Biomolecular network motif counting and discovery by color coding," *Bioinformatics*, vol. 24, no. 13, pp. i241–i249, 2008.

[51] T. K. Saha and M. Al Hasan, "Finding network motifs using MCMC sampling," in *Complex Networks VI*. Cham, Switzerland: Springer, 2015, pp. 13–24.

[52] A. Andersen, W. Kim, and M. Fukuda, "Mass-based nemoprofile construction for an efficient network motif search," in *Proc. IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Social Comput. Netw. (SocialCom), Sustain. Comput. Commun. (SustainCom)(BDCloud-SocialCom-SustainCom)*, Oct. 2016, pp. 601–606.

[53] Y. Hulovatyy, H. Chen, and T. Milenkovic, "Exploring the structure and function of temporal networks with dynamic graphlets," *Bioinformatics*, vol. 32, no. 15, p. 2402, 2016.

[54] B. Schiller, S. Jager, K. Hamacher, and T. Strufe, "Stream-a stream-based algorithm for counting motifs in dynamic graphs," in *Proc. Int. Conf. Algorithms Comput. Biol.* Cham, Switzerland: Springer, 2015, pp. 53–67.

[55] A. Todor, A. Dobra, and T. Kahveci, "Counting motifs in probabilistic biological networks," in *Proc. 6th ACM Conf. Bioinf., Comput. Biol. Health Informat.*, 2015, pp. 116–125.

[56] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2785–2797, Apr. 2015.

[57] P. Audenaert, T. Van Parys, F. Brondel, M. Pickavet, P. Demeester, Y. Van de Peer, and T. Michoel, "Cyclus3D: A cytoscape plugin for clustering network motifs in integrated networks," *Bioinformatics*, vol. 27, no. 11, pp. 1587–1588, 2011.

[58] S. Prakash, H. Agarwal, U. Agarwal, P. Biswas, and S. D. Jaypee, "Discovering motifs in DNA sequences: A suffix tree based approach," in *Proc. IEEE 8th Int. Advance Comput. Conf. (IACC)*, Dec. 2018, pp. 327–332.

[59] D. Tsirogiannis and N. Koudas, "Suffix tree construction algorithms on modern hardware," in *Proc. 13th Int. Conf. Extending Database Technol.*, 2010, pp. 263–274.

[60] G. Han and H. Sethu, "Waddling random walk: Fast and accurate mining of motif statistics in large graphs," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 181–190.

[61] Y. Zhang, B. Wu, Y. Liu, and J. Lv, "Local community detection based on network motifs," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 716–727, 2019.

[62] L. Parida, "Discovering topological motifs using a compact notation," *J. Comput. Biol.*, vol. 14, no. 3, pp. 300–323, 2007.

[63] B. Xu, Y. Liu, C. Lin, J. Dong, X. Liu, and Z. He, "Reconstruction of the protein-protein interaction network for protein complexes identification by walking on the protein pair fingerprints similarity network," *Frontiers Genet.*, vol. 9, pp. 1–10, Jul. 2018.

[64] W. Kim, M. Li, J. Wang, and Y. Pan, "Biological network motif detection and evaluation," *BMC Syst. Biol.*, vol. 5, no. 3, p. S5, 2011.

[65] J. Meier and M. Märtens, A. Hillebrand, P. Tewarie, and P. Van Mieghem, "Motif-based analysis of effective connectivity in brain networks," in *Proc. Int. Workshop Complex Netw. Appl.* Cham, Switzerland: Springer, 2016, pp. 685–696.

[66] O. Sporns and R. Kötter, "Motifs in brain networks," *PLoS Biol.*, vol. 2, no. 11, p. e369, 2004.

[67] E. J. Friedman, K. Young, G. Tremper, J. Liang, A. S. Landsberg, N. Schuff, and A. D. N. Initiative, "Directed network motifs in alzheimer's disease and mild cognitive impairment," *PLoS ONE*, vol. 10, no. 4, 2015, Art. no. e0124453.

[68] M. Middendorf, A. Kundaje, M. Shah, Y. Freund, C. H. Wiggins, and C. Leslie, "Motif discovery through predictive modeling of gene regulation," in *Proc. Annu. Int. Conf. Res. Comput. Mol. Biol.* Berlin, Germany: Springer, 2005, pp. 538–552.

[69] T. Squartini and D. Garlaschelli, "Analytical maximum-likelihood method to detect patterns in real networks," *New J. Phys.*, vol. 13, no. 8, 2011, Art. no. 083001.

[70] G. Fagiolo, J. Reyes, and S. Schiavo, "The evolution of the world trade Web: A weighted-network analysis," *J. Evol. Econ.*, vol. 20, no. 4, pp. 479–514, 2010.

[71] J. He and M. W. Deem, "Structure and response in the world trade network," *Phys. Rev. Lett.*, vol. 105, no. 19, 2010, Art. no. 198701.

[72] F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini, "Randomizing bipartite networks: The case of the world trade Web," *Sci. Rep.*, vol. 5, Jun. 2015, Art. no. 10595.

[73] C. V. Wright, F. Monrose, and G. M. Masson, "On inferring application protocol behaviors in encrypted network traffic," *J. Mach. Learn. Res.*, vol. 7, pp. 2745–2769, Dec. 2006.

[74] L. Liu, C. Han, and W. Xu, "Evolutionary analysis of the collaboration networks within national quality award projects of china," *Int. J. Project Manage.*, vol. 33, no. 3, pp. 599–609, 2015.

[75] Y. Jin, Y. Wei, C. Xiu, W. Song, and K. Yang, "Study on structural characteristics of china's passenger airline network based on network motifs analysis," *Sustainability*, vol. 11, no. 9, p. 2484, 2019.

[76] L. Ping, X. Xing, Q. Zhong-Liang, Y. Gang-Qiang, S. Xing, and W. Bing-Hong, "Topological properties of urban public traffic networks in chinese top-ten biggest cities," *Chin. Phys. Lett.*, vol. 23, no. 12, p. 3384, 2006.

[77] P. Holme and J. Saramäki, "Temporal networks as a modeling framework," in *Temporal Networks*. Berlin, Germany: Springer, 2013, pp. 1–14.

[78] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns, "Network structure of cerebral cortex shapes functional connectivity on multiple time scales," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 24, pp. 10240–10245, 2007.

[79] K. Buza and L. Schmidt-Thieme, "Motif-based classification of time series with Bayesian networks and SVMs," in *Advances in Data Analysis, Data Handling and Business Intelligence*. Berlin, Germany: Springer, 2009, pp. 105–114.

[80] A. Masoudi-Nejad, F. Schreiber, and Z. R. M. Kashani, "Building blocks of biological networks: A review on major network motif discovery algorithms," *IET Syst. Biol.*, vol. 6, no. 5, pp. 164–174, 2012.

[81] R. A. Rossi, N. K. Ahmed, and E. Koh, "Higher-order network representation learning," in *Proc. Int. World Wide Web Conf. Companion Web Conf. Steering Committee*, 2018, pp. 3–4.

[82] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 891–900.

[83] S. Pal, Y. Dong, B. Thapa, N. V. Chawla, A. Swami, and R. Ramanathan, "Deep learning for network analysis: Problems, approaches and challenges," in *Proc. MILCOM IEEE Military Commun. Conf.*, Nov. 2016, pp. 588–593.
[84] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert, "Distance metric learning using graph convolutional networks: Application to functional brain networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 469–477.
[85] F. Monti, K. Otness, and M. M. Bronstein, "MotifNet: A motif-based graph convolutional network for directed graphs," in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2018, pp. 225–228.
[86] S. Mahony and P. V. Benos, "Stamp: A Web tool for exploring dna-binding motif similarities," *Nucleic Acids Res.*, vol. 35, pp. W253–W258, 2007.

SHUO YU received the B.Sc. and M.Sc. degrees from the Shenyang University of Technology, Shenyang, China. She is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian, China. Her research interests include network science, science of scientific team science, and computational social science.

JIN XU received the B.Sc. degree from Dalian Maritime University, Dalian, China, in 2018. He is currently pursuing the master's degree in software engineering with the Dalian University of Technology, Dalian. His research interests include network science and deep learning.

CHEN ZHANG received the B.Sc. degree from the North University of China, Taiyuan, China, in 2019. She is currently pursuing the master's degree in software engineering with the Dalian University of Technology, Dalian, China. Her research interests include network science, data science, and computational social science.

FENG XIA (M'07–SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor with the School of Science, Engineering, and Information Technology, Federation University Australia, and on leave from the School of Software, Dalian University of Technology, China, where he is also a Full Professor. He has published two books and more than 200 scientific articles in international journals and conferences. His research interests include data science, big data, knowledge engineering, social computing, and systems engineering. He is a Senior Member of ACM.

ZAFER ALMAKHADMEH received the M.Sc. and Ph.D. degrees from the Department of Computer Engineering, Faculty of Information and Computer Engineering, Kharkov National Technical University of Ukraine, in 1998 and 2001, respectively. He is currently an Associate Professor with the Computer Science Department, Community College, King Saud University, Saudi Arabia. His main research interests include cloud computing, social network analysis, big data, and intelligent systems.

AMR TOLBA received the M.Sc. and Ph.D. degrees from the Faculty of Science, Menoufia University, Egypt, in 2002 and 2006, respectively. He is currently an Associate Professor with the Faculty of Science, Menoufia University. He is also on leave from Menoufia University to the Computer Science Department, Community College, King Saud University, Saudi Arabia. His main research interests include socially-aware network, the Internet of Things, intelligent systems, big data, recommender systems, and cloud computing.