This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication May 14, 2019, date of current version May 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

A Survey of Measures for Network Motifs

FENG XIA[1], (Senior Member, IEEE), HAORAN WEI[1], SHUO YU[1], DA ZHANG[2], AND BO XU[1]
[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China
[2]Department of Electrical and Computer Engineering, University of Miami, 1251 Memorial Drive, Coral Gables FL 33146, USA

Corresponding author: Bo Xu (e-mail: boxu@dlut.edu.cn).

ABSTRACT Network motifs provide an enlightening insight into uncovering the structural design principles of complex networks across multifarious disciplines, such as physics, biology, social science, engineering, and military science. Measures for network motifs play an indispensable role in the procedures of motif measurement and evaluation which are crucial steps in motif detection, counting, clustering, etc. However, there is a relatively small body of literature concerned with measures for network motifs. In this paper, we review the measures for network motifs in two categories: structural measures and statistical measures. The application scenarios for each measure and the distinctions of measures in similar scenarios are also summarized. We also conclude the challenges for using these measures and put forward some future directions on this topic. Overall, the objective of this survey is to provide an overview of motif measures, which is anticipated to shed light on the theory and practice of complex networks.

INDEX TERMS Network motif, motif measure, network science and motif definition.

I. INTRODUCTION

MINING information hidden in large-scale complex networks has been a hot topic for decades [1], [2]. Most previous works dig up information on basic network structures, such as average degree [3], clustering coefficient [4], average shortest distance [5], betweenness [6], etc. However, previous works have not dealt with mining meso-level network structure information. In 2002, Milo *et al.* [7] defined the concept of "Network Motifs" by frequent connection patterns that significantly exceeds the connection patterns in the randomized networks. Though the concept of motif stems from biological networks, motifs exist widely in various other kinds of complex networks. Motifs have now drawn academic attentions in multiple research areas as well. Motif discovery, motif analysis, motif counting, and other motif-based issues have become the most prevailing research hotspots [8]–[10]. There are fruitful results of its applications in bioinformatics, social statistics, data science, and many other disciplines [11]–[13]. Various kinds of motifs have been found in different scenarios such as gene transcription regulation, food chain, electronic circuit and world wide web [14], [15].

The traditional network properties, which describe nodes, links or the whole network, lack the ability to profile motif structural measures. As it is known, motif structures are significantly different in different networks. Based on this fact, scholars study measures of motifs from the viewpoint of network science. Barabasi *et al.* [16] study motifs and find that motifs play an important role in real world network dynamics. Milo *et al.* [17] propose *Z-score* to evaluate the importance of motifs in networks. Onnela *et al.* [18] study motifs in weighted networks and propose motif-based measures. Abundant research results provide new insights into motif-based network structure studies.

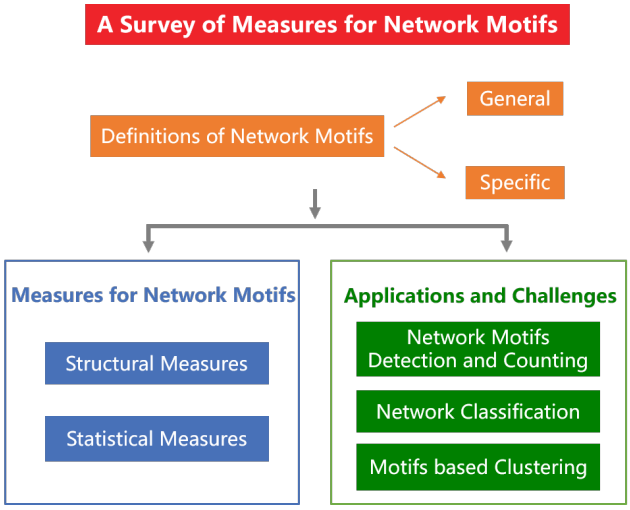

FIGURE 1. The structure of this article.

To enhance the understanding of motifs, scholars inves-

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

tigate from the perspectives of motif application and motif discovery methods, which are obviously important. Rebeiro *et al.* [19] summarize motif discovery algorithms in complex networks. In their survey, they categorize, implement, and analyze motif discovery methods. Other similar surveys mainly focus on studying motifs in biological networks [20]–[22]. Rare literature centralizes motifs in general networks. However, studies about motif measures are fruitful but lack systematicness, which are most included in motif discovery and counting studies. As mentioned above, studies have achieved abundant results of motif structures. Thus, in this paper, we aim at proposing a systematic survey of motif measures. In summary, our contributions are concluded as follows.

- **A systematic survey of motif measures:** In this paper, we summarize almost all motif measures. We classify these measures into two categories, i.e., structural measures and statistical measures. To our best knowledge, we are both the first to introduce motif measures systematically and the first to classify these measures reasonably.
- **Universal measures for network motifs:** We survey universal motif measures, which are fit for general networks instead of biological networks. These measures are described in great detail in this paper.
- **Applications of specific measures:** We introduce motif measures and their different applications, including motif detection, motif counting, network classification, motif-based clustering, etc. Introduction of this content can be a guidance for scholars when choosing motif measures.

The rest of this paper is organized as follows. Section II introduces some basic definitions about motifs. Measures for network motifs are presented in Section III. In Section IV, we introduce applications for network motifs. Conclusions and future topics of research are presented in Section V. The structure of this paper is also shown in Fig. 1. The measures assigned into each category are listed under each heading of subsections respectively in the figure.

II. DEFINITIONS OF NETWORK MOTIF

In this section, we mainly introduce the formal definition of network motifs and briefly review several definitions used in the specific networks. For the sake of simplicity, we do not distinguish the terms "network" and "graph" in the following paper. We model a network as a graph $G = \{V, E\}$, where $V = \{v\}$ is the set of all nodes or vertices, and $E = \{e\} = \{(u, v)\}$ is the set of edges or links. The edge connects a pair of nodes (u, v). $|V|$ denotes the number of nodes in G, which is called the graph size or order. If an edge $e = (u, v) \in E$ is ordered, it is called a directed edge. On the contrary, it is called an undirected edge. If all the edges in a graph are directed, the graph is called a directed graph. If they are all undirected, the graph is an undirected graph. A path between two nodes (u, v) consists of a sequence of nodes that starts from node u and ends with node v. Each of nodes in the sequence is adjacent to its successor and predecessor. An undirected graph with no multiple edges is a completed graph if there are $n(n-1)/2$ edges in it.

Let $G_k \subset G$ or k-subgraph be the subgraph of G whose size is k. An induced subgraph means that it need to include all edges connected to the nodes existing in the original graph. The neighborhood set $N(v)$ of node $v \in V$ consists of nodes connected to v. Two graphs G and G' are isomorphic if there is a bijection function $f_{ismp} : V' \to V$ with $(u, v) \in E' \Leftrightarrow (f_{ismp}(u), f_{ismp}(v)) \in E$ for all $u \in V'$, $v \in V'$. Here f_{ismp} is called an isomorphism between graph G and G'.

A. NETWORK MOTIF DEFINITIONS

Milo *et al.* [7] first propose the definition of "network motifs" as patterns of inter-connections occurring in complex networks which occurs much more frequently in the original network than in the similar randomized networks. In this definition, motifs are statistically over-represented so that in the randomized networks the in and out degrees (as degrees in undirected graphs) of all single nodes should be equal to those in the original networks [23]. When calculating the significance of G_k, the number of all G_{k-1} appearing in randomized networks should be equal to that in the real network. Based on that, Riberio *et al.* [19] give a more formal version of the definition that is described as follows:

Given a network, a set of parameters $\{P, U, D, N\}$ and an ensemble of N similar networks, network motif is defined as an induced subgraph appearing in the real network when it satisfies the following conditions:

1) $p((\bar{f}_{rand}(G_k) > f_{real}(G_k))) \leqslant P$
2) $f_{real}(G_k) \geqslant U$
3) $f_{real}(G_k) - \bar{f}_{rand}(G_k) > D \times \bar{f}_{rand}(G_k)$

where $f_{real}(G_k)$ is the frequency of a motif in the real network and $\bar{f}_{rand}(G_k)$ is the average frequency in all randomized networks. The parameter P is a probability threshold determined by an ensemble of a large number of N similar randomized networks and the *Z-score* measure that we will introduce in III. U is an uniqueness cutoff value for the frequency of a motif in the real network and D is the proportional cutoff to guarantee the minimum difference between $f_{real}(G_k)$ and $\bar{f}_{rand}(G_k)$.

The first constraint means the probability of the frequency of a motif in randomized networks is greater than that in the real networks. Meanwhile, the frequency in both randomized networks and real networks should be lower than the threshold P, which is to ensure that the motif is over-represented. It is called *P-value* as a threshold measure for network motifs, which could be represented as

$$P_{value} = \frac{1}{N} \sum_{i=1}^{N} \delta(c(i)), \quad c(i) : f_{rand}(i) \geqslant f_{real}(i) \quad (1)$$

where N still denotes the number of randomized networks. The value of function $\delta(c(i))$, i.e., Kronecker delta function, is 1 when the condition $c(i)$ holds [24]. The second constraint

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

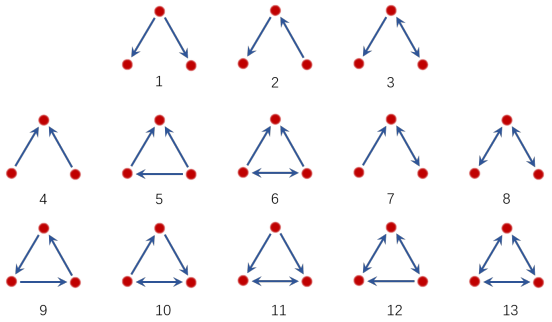

FIGURE 2. All possible directed 3-motifs.

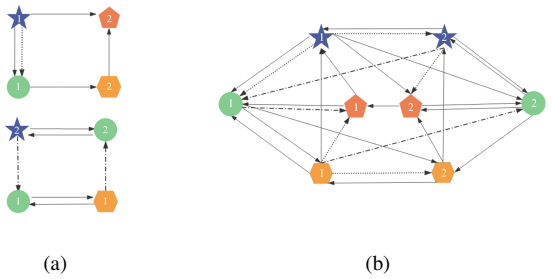

FIGURE 3. Simple military communication network and supernetwork motifs.

guarantee the lower bound of the frequency of a motif in the real network. The last one makes the number of appearances in the real network be significantly larger than that in the randomized networks. This is to ensure the frequency of motifs with consistency in randomized networks and the real network. Moreover, these motifs are not under a narrow distribution in a randomized network.

In Milo's experiments, the parameters $\{P, U, D, N\}$ are the set as $\{0.01, 4, 0.1, 1000\}$ [7]. Whereas for various aims or in different networks, we may adjust the values of these parameters. This formal definition has been properly used in directed and undirected networks with various sizes in many fields, which is a fundamental definition in correlated research. But there are certain limitations that we will ignore the patterns or subgraphs with more important functions but less statistical significance.

In Fig. 2 we exhibit all the thirteen possible connected motifs as examples. The majority of research considers only connected subgraphs as possible motifs. Whereas some studies also analyze unconnected subgraphs [25]. Here in this paper we refer to connected subgraphs unless specially stated.

The definition can also be explained from the statistical perspective. For the original network and its randomized networks whose size and degree distribution are the same, they have a common feature as the original network where the feature refers to the frequency of one or several kinds of subgraphs. Here the ensemble of randomized networks is regarded as the null model. The three constraints are the validation criteria that should hold simultaneously, which makes the null hypothesis above invalid. Except for null model, Markov chain algorithms can also be used for randomization of networks [26]. In the definition based on null model, the distribution of features in the underlying network is considered as Gaussian. However some researchers suggest that under null model or Gaussianity is not sufficient to assess its significance. Other distributions i.e., compound Poisson distributions are used to determined whether the motif is overrepresented [27]. Kernel density estimation and cross validation are also used to learn the distribution [28].

There is also an alternative definition, Berg *et al.* perform a subgraph alignment across an ensemble of small graphs to extract an average structure named consensus pattern [29]. This definition seems to reduce the effect of data incompleteness, but the process of alignment among many different subgraphs may lose some information. And overlaps of motifs are also considered in the definition in some researches [28], [30], [31].

B. SPECIFIC DEFINITIONS

In specific application contexts, network motifs have been studied from multiple perspectives. In this section, we introduce some correlative definitions of network motif deformations and the motif definitions in specific networks.

Anti-motif is the statistical insignificance subgraph, opposite to the aforementioned definition of motif, which also makes sense to existing researches. It is suggested that anti-motif should satisfy: $p(\bar{f}_{rand} < f_{real}) < P$ and $\bar{f}_{rand} - f_{real} > D \times \bar{f}_{rand}$ [7]. It means the frequency of anti-motif is lower than the expected value according to the randomized networks [32].

Maximal motif is another essential concept characterizing the maximality of motifs [33], [34]. If one motif is not contained in any other motif of G, that motif is considered as a maximal motif. Only detecting maximal motifs is of benefit to reduce computations because it costs more storage to detect the subgraphs than all motifs.

In biological networks, network motifs are supposed to be recurring circuit elements with key information processing tasks [21]. Network motifs have also been applied to the supernetworks. Supernetwork can be easily comprehended as a composite network of various subnetworks. Shi *et al.* [35] summarized the definition of supernetwork as a multi-linked heterogeneous network including various links and nodes, which focuses on the entire network function. Fig. 3 presents a simple military communication supernetwork (MCSN) and two supernetwork motifs are evolving from it. Fig. 3(b) is a simple military communication supernetwork. The nodes in shape of circle, star, pentagon and hexagon denote decider, sensor, target and influencer, respectively. The solid line denotes situation message transmission. The dotted line denotes command and control (C2) message transmission. The dashed line denotes state message transmission. In Fig. 3(a), there are two examples of supernetwork *4-motif*s that can be found in Fig 3(b). *k(m)-motif* denotes the supernetwork motif

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

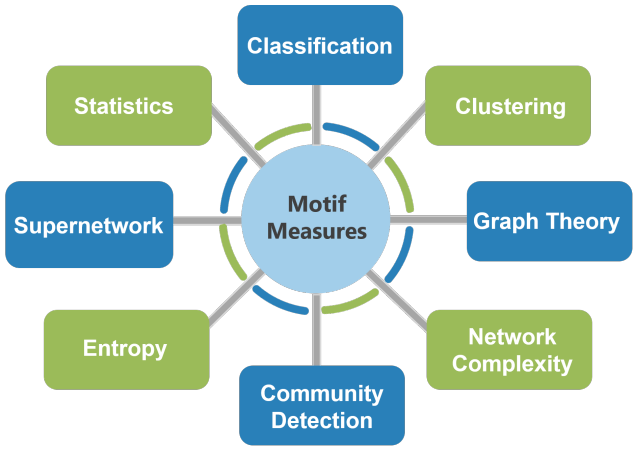

FIGURE 4. Subjects and techniques related to motif measures. The definition of motif measures is related to statistics, graph theory, clustering, classification, network complexity, entropy, supernetwork and community detection.

composed of k nodes with m types [36]. In supernetwork motif the node with maximal degree is regarded as a core node named Core of Supernetwork Motif. If there are more than one node having the maximal degree in a supernetwork motif, the motif is called multi-core motif. If all nodes in the motif have the same degrees, the motif is called *np-core* motif. In processing complex supernetwork analysis, we can analyse *n-motif* as a structural motif. For further analysis of function, *n-motif* can be partitioned into *n(k)-motif* as a functional motif.

III. MEASURES FOR NETWORK MOTIF

Studies on measures for network motif combine not only knowledge of graph theory, statistics, supernetwork, entropy, network complexity and community detection, but also techniques of machine learning like clustering and classification. The relations of measuring network motifs are shown in Fig. 4, whose basic information is the basis of the following comprehension. The measures for network motifs can be divided into two categories, structural measures and statistical measures. Structural measures describe the topological features networks from microcosmic view like the significance of node and edge, or from macroscopic view for example the complexity of network. Statistical measures describe the statistical significance of a motif or other features on the basis of motif frequency or motif statistical significance. The classification measures are shown in Fig. 5, in which the blue parts are structural measures and the green parts are statistical measures. In this section, we specifically introduce each measure and its applications.

A. STRUCTURAL MEASURES

Structural measures make sense for network motifs. For example, randomized networks can be constructed with the same topological measures, size and degree distribution, as the original network for motif detection. In this section we

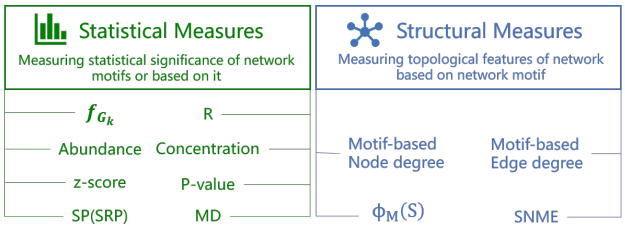

FIGURE 5. Motif measures are divided into two categories, i.e., structural measures and statistical measures. Their definitions are written in the green and blue hexagons, respectively.

firstly introduce structural properties in graph theory. Then we also introduce some other topological structure measures for network motif.

TABLE 1 lists all possible motifs of 3 and 4 nodes in undirected graphs. We also list the basic measures inherited from graph theory, such as density ρ, max degree Δ, mean degree $\bar{d}$, assortativity γ, total number of triangles $|T|$, max k-core number K, Chromatic number χ, diameter D, the max betweenness B, the number of components $|C|$.

TABLE 1. Structural Properties for Undirected 3&4-Motifs.

	Connection Patterns							
Motif								
ρ	1.00	0.83	0.67	0.67	0.50	0.50	1.00	0.67
Δ	3	3	3	2	3	2	2	2
$\bar{d}$	3.0	2.5	2.0	2.0	1.5	1.5	2.0	1.33
γ	1.00	-0.66	-0.71	1.00	-1.00	-0.50	1.00	-1.00
$\|T\|$	4	2	1	0	0	0	1	0
K	3	2	2	2	1	1	2	1
χ	4	3	3	2	2	2	3	2
D	1	2	2	2	2	3	1	2
B	0	1	2	1	3	2	0	1
$\|C\|$	1	1	1	1	1	1	1	1

1) Motif-based Node Degree

Motif-based node degree, presented by Han *et al.* [37], is a measure combining node degree and network motif. Given a network G and the motif M of G, Motif-based node degree is written as d_i^M, which denotes the number of Ms including node i. Generally, an edge connecting two nodes can be seen as a *2-motif*. Then for node i, the traditional node degree can be described as the number of *2-motifs* that include node i, which is a special case of the motif-based node degree.

2) Motif-based Edge Degree

Motif-based edge degree is also proposed by Han *et al.* [37]. There has been measures for edge-clustering coefficient, which is the ratio of the actual number of triangles including the edge e to the number of all possible triangles including e. For an edge connecting two nodes i, j, the formula of edge-

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

clustering coefficient, $C_{ij}^{(3)}$ is

$$C_{ij}^{(3)} = \frac{z_{ij}^{(3)} + 1}{\min[d_i - 1, d_j - 1]}, \tag{2}$$

where $z_{ij}^{(3)}$ is the number of triangles formed by the edge. But it does not consider that in different networks, there are different kinds of motifs. For example, there is no *3-motif* in hierarchical tree structure networks. So, given an arbitrary network G and its motif M, for an edge e, the edge degree based on motif is redefined as the number of Ms including e, denoted by C_e^M.

The motif-based node and edge degree are analysed in terms of Karate network, Dolphin network, Newman science coauthors network, PGP network and email network. The authors calculated the Pearson correlation coefficient between the degree they presented and traditional concepts [38]. It is represented as r and the formula is shown in (3).

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}} \tag{3}$$

In experiment results, r is high enough to verify the rationality and superiority of the motif-based node degree and edge degree.

3) SNME

Supernetwork Motif Entropy (SNME) is a complexity measure combining the heterogeneity of nodes, the significance of functional motifs and information entropy for military communication networks (MCN). It is presented by Shi *et al.* [36]. The mathematical representation is

$$SNME = \sum_{i=1}^{m'} P_i \sum_{j=1}^{m'} H_i^{(m'(j))}, \tag{4}$$

wherein m' denotes the number of node types in the whole network. P_i denotes the proportion of the node number belonging to type i in the network. $H_i^{(k(j))}$ denotes the $m'(j)$-*motif* entropy whose core node is included in type i. The *k(m)-motif* entropy, $H^{(k(m))}$, is defined by information entropy as

$$H^{(k(m))} = -\sum_{i=1}^{M_{k(m)}} p_i^{(k(m))} \log_2 p_i^{(k(m))}, \tag{5}$$

where $p_i^{(k(m))}$ is a probability distribution of *k(m)-motif* that belongs to type i. $M_{k(m)}$ means the number of types that *k(m)-motif* can formate. If $m = 1$, then $H^{(k(m))} = H^{(k)}$, which indicates the *k(m)-motif* entropy without considering the heterogeneity of nodes and applicable for motifs in common networks [39].

According to other studies about network complexity and Boltzmann's entropic equation, the number of states of network possible structures will increase with the number of node types. So *SNME* can be used as a complexity measure for MCNs. More concretely, *SNME* can measure the complexity of network from the perspective of classification of nodes' function, combining the existing measures for network structural complexity, such as *OdC* [40], *MAg* [41], *Cr* [42] and *Orb* [43]. For different aims, we can choose different size of supernetwork motifs, which will give a new insight into MCNs analysis. But the effect of it is not good enough when measuring network complexity without other measures. Its application in networks of other domains has not been studied.

4) $\phi_M(S)$

$\phi_M(S)$ is a structural measure for evaluation clustering or partition of network motifs [11]. It is a generalization of conductance metric that is a useful graph partitioning score in spectral graph theory [44]. The measure can be calculated as the following equation

$$\phi_M(S) = \frac{cut_M(S, \bar{S})}{\min[vol_M(S), vol_M(\bar{S})]}, \tag{6}$$

where S denotes a set of nodes partitioned in a cluster, and $\bar{S}$ is the complement of $S \subset V$. $cut_M(S, \bar{S})$ is the number of motif Ms that has at least one node in S and one node in $\bar{S}$. $vol_M(S)$ denotes the number of nodes belonging to motif Ms and moreover residing in S. This measure reflects the extent of the motif conductance of cluster S relative to M. The smaller the value is, the better the cluster is.

B. STATISTICAL MEASURES

As described in Section II, statistical significance, especially compared with randomized networks, is one of the most important features for network motifs. Several measures related to statistics of motifs have been proposed for diverse purposes, such as discovery of motifs and classification of networks. In this section, we review the statistical measures for network motifs arranged from simple to complex. For each measure, we summary its application contexts.

1) Frequency

As mentioned above in definitions, the frequency of a motif f means the number that a motif appears in a network. It is a necessary condition for a subgraph that the frequency of it in real network is higher than the mean frequency in similar randomized networks. Notice that the overlap allowed can make a big difference on both the value of frequency and the counting complexity [21], [45]. Thus it is indeed important to distinguish the frequency when detecting motifs. There are three conditions where they allow arbitrary overlaps of nodes and edges, only overlaps of nodes and no overlaps [45]. In biological networks, for example PPI (protein-protein interaction) networks, one node may take part in different motifs that denotes modules performing different functions. In that case, overlaps of nodes or edges are considered as the first condition. On the contrary, no overlaps of nodes nor edges are considered in algorithms presented in data mining papers [46].

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

Motif frequency has a wide application in different networks. Marinho *et al.* [47] study the authorship attribution using the frequency of motifs. In the word co-occurrence networks derived from written texts, they calculate the absolute frequency of all thirteen directed 3-motifs as the only feature of classification and find that there is a dependency between the frequency and the writing style of different authors.

We can also calculate a ratio of the frequency in real network to the mean frequency in randomized networks as $R = f_{real}/\bar{f}_{rand}$. It should be higher than 1.1 generally. There are some softwares that can extract the frequency of motifs, such as *mfinder* and *fanmod* [41].

2) Concentration

Concentration is also an interesting measure for network motifs in combination with sampling methods. Given a subgraph G_{ki}, the concentration is defined as

$$C(G_{ki}) = \frac{f(G_{ki})}{\sum_j^{card(G_k)} N(G_{kj})}, \tag{7}$$

where $f(G_{ki})$ refers to the frequency of one specific subgraph in size k, and $\sum_j^{card(G_k)} N(G_{kj})$ is the total number that all possible subgraphs in the same size k appearance in the network. This measure is also used in experiments when comparing motifs from PPI network and Internet routers network [7].

3) Z-score

For each subgraph of G_ks, the statistical significance compared to that in randomized networks is qualitatively described as the measure named *Z-score* [7],

$$Z = \frac{f_{real} - \bar{f}_{rand}}{std(f_{rand})}, \tag{8}$$

where f_{real} is the frequency or the number that G_k appears in the real network. $\bar{f}_{rand}$ is the average that G_k appears in the randomized networks. $std(f_{rand})$ is the standard deviation that G_k appears in randomized network ensemble. *Z-score* is widely used in network motif discovery and counting tasks [25]. Whereas it asserts the samples, i.e., subgraphs, are in Gaussian distribution. But in real-world networks, subgraphs are often not Gaussianity [28]. *Z-score* tends to be higher for motifs in respective large networks. So other measures of statistical significance for motifs are proposed.

4) Abundance

The abundance is another measure of the characterization of motif statistical significance written as $\triangle$ [17], [21]. Similar to *Z-score*, for each subgraph, abundance measure defined as

$$\triangle = \frac{f_{real} - \bar{f}_{rand}}{f_{real} + \bar{f}_{rand} + \epsilon}, \tag{9}$$

where ϵ is a value to ensure the absolute value of $\triangle$, $\mid \triangle \mid$ will not be too large to mislead the results when the subgraph occurs very few times in both the original and randomized networks.

5) SP

Milo *et al.* [17] proposed a relative significance measure, the significance profile (SP), for network motifs statistical significance compared to randomized networks. The SP is a normalized value of *Z-score* vector to length one, defined as (10)

$$SP = \frac{Z_{G_{ki}}}{(\sum^{card(G_k)} Z^2_{G_{kj}})^{1/2}}, \tag{10}$$

where G_{ki} denotes a subgraph i in size of k, for each G_{ki}, $Z_{G_{ki}}$ is its *Z-score* value. SP can effectively avoid the impact of different sizes of networks.

Milo *et al.* also propose a specific measure as Triad Significance Profile (TSP) for 3-motifs. They analyze all 13 directed and connected subgraphs' TSP for networks that were constructed from different domains. They find that some networks have similar TSPs whose correlation coefficients are higher than 0.99, which are divided into several groups named as superfamilies. For example, the sensory transcription networks from the bacteria Escherichia coli, Bacillus subtili and the yeast Saccharomyces cerevisiae have similar TSPs are regarded as one superfamily with rate limited feature. In these networks, the expected response times are as short as that of the components in the network, i.e., motifs. Furthermore, WWW networks and social networks have similar TSP so that they can be recognized as a superfamily, which indicates that classical models of social structural organization may be benefit to the comprehension of WWW structure. We can also calculate the correlation between the TSPs of different networks which also can be used to cluster networks as distinct superfamilies.

TSP is robust to data under missing circumstance or with random errors. Experiments demonstrate that the TSP is almost insensitive to removing 30% of the edges or adding 50% new edges randomly. But TSP can be sensitive to rare occasions of a mutual edge in a network.

For 4-motifs in both directed and undirected networks, the normalized *Z-score*, SP, can be seriously affected by the size of network. Instead of *Z-score*, the normalized abundance is used to describe the statistical significance of 4-motifs. Thus subgraph ration profile (SRP) is proposed and its definition is

$$SRP_i = \frac{\triangle_{G_{ki}}}{(\sum^{card(G_k)} \triangle^2_{G_{kj}})^{1/2}}. \tag{11}$$

This measure is calculated in electrical power grid networks, protein structure networks and AS (Autonomous systems) networks, measuring its capability of classification of different networks. However, it is interesting that some networks of different types in the same TSP superfamily present diverse SRPs, measuring that higher order subgraph profiles can promote the network classification.

6) MD

Motif Difficulty (MD) is a problem difficulty measure for evolutionary algorithms (EAs) as presented in [48]. It can quantify the difficulty of different problems into $[-1, 1]$,

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

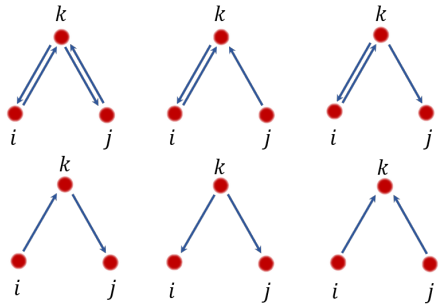

FIGURE 6. Basic motifs. These six motifs are subsets of other directed 3-motifs

where $MD = -1$ means the problem is the easiest, and $MD = 1$ means it is the most difficult. Since it is based on the statistics of motifs and nodes in motifs, we classify it as statistical measures. Before describing the formula of MD, the directed fitness landscape network and three types of motifs used are introduced.

A directed fitness landscape network is defined as $\overrightarrow{FLN} = (V, \overrightarrow{E})$. Here V is the ensemble of nodes in the network, and each node denotes a configuration in the search space or a point in the fitness landscape. $\overrightarrow{E}$ is the set of edges and each edge connects a certain node to one of its neighbours. The direction of the edge is determined by the value of the fitness function f. The edge (x, y) points to y from x when $f(y) \geqslant f(x)$.

As shown in Fig. 6, the authors choose six types of 3-motifs that are the subsets of other 3-motifs, called basic motifs. Unlike other measures that define motifs based on comparison with that in randomized networks, here we choose the motifs because they are the subsets of other motifs, since generation randomized networks of FLN is another research topic with large difficulties. Then a distance motif is defined on basic motifs as $M^d = \{V_M, \overrightarrow{E_M}, d_1, d_2, d_3\}$, where $d_{i,i=1,2,3}$ refers to the Hamming distance between d_i and the global optima. The distance and fitness information are both considered when checking the searching direction in a motif. So effective path is presented as a path in M^d with no inverse direction edges existing for each edge in the path. If there is no effective paths in M^d, it is a neutral motif M^N. If all effective paths in M^d satisfy $d_{start} \geqslant d_{end}$, M^d is a guide motif M^G. In all other cases, it is named as deceptive motif M^D. Based on the description above, a network level difficulty measure MD^e has been presented and the formula is:

$$MD^e = \frac{\sum_{i=1}^{|V|-2}\sum_{j=i+1}^{|V|-1}\sum_{k=j+1}^{|V|}(|M^D_{i,j,k}| - |M^G_{i,j,k}|)}{\sum_{i=1}^{|V|-2}\sum_{j=i+1}^{|V|-1}\sum_{k=j+1}^{|V|}|M^d_{i,j,k}|} \tag{12}$$

where $i, j, k \in V$. $|M^D_{i,j,k}|$, $|M^G_{i,j,k}|$, and $|M^d_{i,j,k}|$ represent the number of deceptive, guide and distance motifs that any three nodes i, j, k form, respectively.

The value of MD^e is in the range of $[-1.0, 1.0]$. When it is -1.0, i.e., all M^ds in DFLN are guide motifs, the problem is the easiest. When MD^e is 1.0, all M^ds are deceptive motifs, so that the problem is the most difficult. The measure MD^e predicts problem difficulty in the most straightforward way by counting the numbers of distance motifs over the whole network. So it cannot reflect the difference in details.

The measure MD^o is presented to describe the detailed feature and spatial distribution of each kind of distance motifs. In an M^G, there is an effective path whose length is two and the distances decrease from the start node to the end node in turn. Then this M^G is a *Core Guide Motif*. The start node with the smallest distance is the *Core Guide Node*. Similarly, in an M^D if there is a path of length two and the distance increase from start to end in this path in turn, it is a *Core Deceptive Motif*. The start node is called *Core Deceptive Node*. Based on the core guide node and core deceptive node, another measure named node level difficulty is defined as

$$MD^o = \frac{|(V - V_{CG}) \cap V_{CD}| - |V_{CG}|}{|V|} \tag{13}$$

where $(V - V_{CG}) \cap V_{CD}$ refers to the set of core deceptive nodes that are not the core guide nodes. The value is also range from -1.0 to 1.0 This measure considers the situation of each node. In searching process, guide motifs always have priorities. Even if a node belongs to both guide and deceptive motifs, the deceptive motif may not be visited in the searching process. So only core guide nodes and the core deceptive nodes who are not simultaneously core guide nodes are taken into account.

Combining the general and detailed measure for problem difficulty, the authors integrated MD^e and MD^O and proposed a measure *Motif Difficulty* labeled as MD.

$$MD = \frac{MD^e + MD^o}{2} \tag{14}$$

Obviously, the value of MD also ranged in $[-1.0, 1.0]$. For the easiest problems MD is -1.0 and for the most difficult problems MD is 1.0. Since it is computational costly and impractical to compute on the whole network quickly, an approximate measure may be preferred. A sampling technique has been proposed by computing on a sample of the search space. It is verified by experiments that the approximate MD is stable [48]. It is validated that MD is not only consistent with other difficult measures on previous results of problem difficulty, but also performs well on some counterexamples for other difficulty motifs. It has the advantages of being insensitive to nonlinear scaling, detecting the presence of constantness and being robust to irrelevant deceptive information. But there are also some disadvantages. It takes global optima as reference points that are impractical to find. The operator used to build FLNs has constraints on MD. The distance is calculated by Hamming distance instead of the ideal distance named the shortest path length.

We give a brief summary of all measures mentioned above in TABLE 2. We also combine the measures' application networks and target problems except for the category and

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

TABLE 2. Measures for Network Motifs and Application Scenarios

Classification	Measure	Definition	Application network	Target problem
Structural Measure	Motif-based Node Degree	d_i^M	all networks	node significance
	Motif-based Edge Degree	$C_{ij}^{(3)} = \frac{z_{ij}^{(3)}+1}{\min[d_i-1, d_j-1]}$	all networks	edge significance
	SNME	$SNME = \sum_{i=1}^{m'} P_i \sum_{j=1}^{m'} H_i^{(m'(j))}$	supernetworks with heterogeneous nodes	network complexity
	$k(m)-motif$ Entropy	$H^{(k(m))} = -\sum_{i=1}^{M_{k(m)}} p_i^{(k(m))} \log_2 p_i^{(k(m))}$	supernetworks with heterogeneous nodes	motif entropy
	$\Phi_M(S)$	$\Phi_M(S) = \frac{cut_M(S,\bar{S})}{\min[vol_M(S), vol_M(\bar{S})]}$	all networks	evaluation of motif clustering
Statistical Measure	Frequency	f_{G_k}	all networks	motif counting and statistical significance
	R	$R = f_{real}/\bar{f}_{rand}$	real networks with randomized networks	motif statistical significance relative to different sizes of networks
	P-value	$P_{value} = \frac{1}{N}\sum_{i=1}^{N}\delta(c(i))$ $c(i): f_{rand}(i) \geqslant f_{real}(i)$	real networks with randomized networks	motif statistical significance relative to randomized networks
	Concentration	$C(G_{ki}) = \frac{f(G_{ki})}{\sum_j^{card(G_k)} N(G_{kj})}$	noncomplete networks	motif statistical significance relative to that under Gaussian distribution
	Z-score	$Z = \frac{f_{real}-\bar{f}_{rand}}{std(f_{rand})}$	real networks with randomized networks	motif statistical significance relative to randomized networks
	Abundance score	$\Delta = \frac{f_{real}-\bar{f}_{rand}}{f_{real}+\bar{f}_{rand}+\epsilon}$	real networks with randomized networks	motif statistical significance relative to randomized networks
	SP	$SP = \frac{Z_{G_{ki}}}{(\sum^{card(G_k)} Z^2_{G_{kj}})^{1/2}}$	real networks with randomized networks	3-motif statistical significance relative to randomized networks
	SRP	$SRP_i = \frac{\Delta_{G_{ki}}}{(\sum^{card(G_k)} \Delta^2_{G_{kj}})^{1/2}}$	real networks with randomized networks	4-motif statistical significance relative to randomized networks
	MD	$MD = \frac{MD^e+MD^o}{2}$	fitness landscape networks	evolutionary algorithm complexity

definition. Measures applications are further discussed and analysed in the next section.

IV. APPLICATIONS AND CHALLENGES

The measures for network motifs are used for solving problems related to motifs. In this section, we summarize three important problems related to motif and list the measures for each problem. As shown in TABLE 3, we make the summary of the application problems of each measure. *f* and *R* can be used in all applications we listed, which are basic but important measures. In terms of the studies for these problems, we also analyze the facing challenges.

A. NETWORK MOTIF DETECTION AND COUNTING

Network motif detection is also referred to as motif discovery or motif finding [49]–[51]. It is similar to motif counting tasks [25], [52], [53]. It is important and basic problem in studies of network motifs [54].

Motif detection process includes several tasks such as motif definition, the definition of frequency concepts, generation of randomized networks, choosing measures for statistical significance and determining the isomorphism of subgraphs.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

We have discussed these tasks in Section II and III. Here we mainly focus on the measures to be used in these tasks. The measures to be used for detection includes all statistical significance measures. As mentioned in Section III, we can enumerate the measures as: *f, R, P-value, Z-score, Abundance, SP* and *SRP*.

The first challenge is to choose an appropriate measure for the target network. There is a precondition that the frequency of the subgraphs are under Gaussian distribution. But it is not appropriate to all networks. For example, the distribution may be undersampled. So some methods based on other distributions like compound Poisson distribution are applied in motif counting and detection. Kernel density estimation, cross validation can be used for density estimation. But few measures are presented.

In addition, randomized network generation is also a challenge. There are a few definitions from other perspectives or at least without randomized networks. For large complex networks, generation of randomized networks with corresponding features that need to go through the whole network is a task with high time complexity. It is impractical to form a set of randomized networks in some cases, i.e., randomized network construction for fitness landscape networks is also a research topic whose time complexity is high, relatively. Thus in this kind of problems, measures without randomized networks to be used are proper for these problems, like *concertration*, *frequency*.

Functional significance is also important especially for networks with complex function. Here *MD* considers function and statistics of motifs, but it is not used for motif detection. Researchers of biology are interested in predicting the functional behavior of motifs from its structural features and finding the necessity between biological significance and the abundance of such motifs [55].

Another challenge is the time complexity and computational cost. Determining whether subgraphs are isomorphic is a computation-intensive task, which has been proved an NP-hard problem. The computational cost will increase rapidly with network size and motif size. The scale and density of the real world networks are also growing increasingly. So approximate algorithms for distinguishing larger isomorphic subgraphs are in urgent need.

Several strategies have been used for these tasks. One uses subgraph sampling through the network instead of exact enumeration. Another strategy is the motif centric approach, which can reduce isomorphism computational cost together with symmetry breaking and mapping methods [50]. But it is not fit for large motifs because the number of the types of motifs grow exponentially with motif size. Many authors will apply their measures or algorithms for larger motifs in future work [25], [26], [47]. At present, there is no algorithm that can detect motifs having more than 10 nodes from a large and complex network in a practical time.

To sum up, the challenges in network motif detection and counting related to measures can be summarized as three points. Firstly, we need statistical significance measures for different underlying distributions. Secondly, we may investigate about measures reducing the computations and complexity on generation and analysis of randomized networks. Last but not least, we should try to reduce the computational cost in the detection and counting process.

B. NETWORK CLASSIFICATION

Network classification means classifying the networks with different sizes from various domains into several groups by the features of network motifs [17], [47], [56], [57]. In other words, we evaluate the features of subgraphs included in different networks and compare these features.

Milo *et al.* [17] present *SP, SRP* and adapt them in different networks. Then the measures are used in email-based social networks [58], [59]. The measures are also used in text networks [47], time series networks [60], Internet networks, etc. Its core idea is also about statistics of motifs in the network. After calculating the relative statistical significance measures for all possible motifs then all the results of the measures construct the profile. If the profiles of two networks are much similar, they fall into the same group. Using this method, different writing style of authors, sensors with different functions and networks from different domains can be distinguished clearly. The networks that are commonly thought with no similarities may be assigned into the same group by this method. These networks may have some internal similarities that have not been discovered.

One way is to explore the generation process and evolution law of these networks. Another interesting result is that *4-motif* and *3-motif* analysis form distinct classification results. We can not distinguish which k order motif reveals the network profile better. So we can only combine its known real world function unit to analyze the profile, but can not find more unknown function or organization of the networks.

There is a similar task described as comparison of networks with graphlets, which means analyzing the graphlet degree distribution of different networks and the "agreement" to compare their similarity or difference [61]. Notice that graphlet is a concept which is as same as motif in structure, but different in statistical significance. There is no need to compare the frequency of the graphlet to the randomized networks for determining whether it is a graphlet or not. Because of the structural similarity between motif and graphlet, the studies on network classification with graphlet can lead to inspiration to that with the motif.

The challenge we are facing in studying this problem is the analysis of high order motifs. The appropriate region for motif sizes for different size of networks need to be studied. For instance, we can analyze the relationship between motif size and network size when performing classifying networks. The relative value is preferred for a better universality.

C. MOTIF BASED CLUSTERING

Motif based clustering is a hot research topic that has attracted increasing interest [11], [62]–[66]. We perform clustering nodes on a network previously. But the scale of real world

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

TABLE 3. Application Scenarios for Each Measure

Function \ Argument	f	R	*Concentration*	*Abundance*	*P-value*	*Z-score*	*SP(SRP)*	$\Phi_M(S)$
Detection&Counting	✓	✓	✓	✓	✓	✓		
Motif Based Clustering	✓	✓		✓				✓
Network Classification	✓	✓				✓	✓	

networks and the knowledge in them grow explosively, which makes it difficult and computational costly for the tasks like community detection and conflation of hub structure in geography. In addition, triangles are of equally practical significance as well as pairs of nodes that have been analyzed and concluded in social and biological networks. When we perform clustering tasks through networks, it is reasonable to keep the nodes in a motif, such as a triangle, in a cluster. Therefore, clustering based on motifs is studied on the social network, biological network, transportation network, etc. Different from clustering based on nodes, motif clustering aims at reducing cutting both edges and motifs essential to the target problem. Overlapping of community or clusters should be considered as well. When choosing methods or measures for motif clustering, another noteworthy problem is whether it is applicable for the target networks. There are few methods or measures applicable to all directed, undirected, weighted signed networks and hypernetworks.

The measures that can be used in clustering problems except for the statistical significance is $\phi_M(S)$, which is an evaluation measure for clustering. The method is based on spectral clustering. There are also some other clustering strategies such as correlation clustering and embedding methods. The challenges in motif based clustering are also related to the exponential explosion as the size of motif increases. It is also important to choose appropriate clustering strategies. Although the strategy based on spectral clustering has been studied widely, some researchers consider it without general analytical guarantees [64]. And using parallel and low complexity methods can be regarded as a challenge.

Network motifs can be used to study many other problems, such as evaluating the significance of node and edge in a network, measuring the complexity of networks and supernetworks, etc. And the challenges are mostly bound up with the size of motif, computational cost, approximate algorithms and the deviation from the underlying network.

V. CONCLUSION

Network motifs have attracted immense attention over decades due to the enormous applicability in understanding complex networks. Various measures for network motifs have been proposed in the detection and evaluation of motifs in complex networks. In this paper, we present an organized and detailed overview of the state-of-the-art measures for networks motifs and categorize them as structural measures and statistical measures. We study the definition and application scenarios of each measure and showcase the correspondences between measures and application scenarios in tables. We also enumerate the available and appropriate measures for three essential problems with respect to network motifs. Structural measures are applicable to measuring the importance and goodness at meso-level. Statistical measures describe the comparative frequency of motifs in the whole network from various perspectives, which are indispensable measures for motif detection and counting. Furthermore, we summarize the challenges for motif measure computation and choosing. Since the complexity of the networks, it is arduous to circumvent the computational intractability in generation randomized networks and detection of high-order motifs. The deviation between the model and the underlying network makes the choice of measures a formidable task. Consequently, there arises a need for measures with less computational cost and approximate algorithms for the existing measures, which are still open research issues that we intend to investigate.

REFERENCES

[1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[2] X. Su, W. Wang, S. Yu, C. Zhang, T. M. Bekele, and F. Xia, "Can academic conferences promote research collaboration?" in *Digital Libraries*, 2016, pp. 231–232.

[3] D. Antonakaki, S. Ioannidis, and P. Fragopoulou, "Utilizing the average node degree to assess the temporal growth rate of twitter," *Social Network Analysis and Mining*, vol. 8, no. 1, p. 12, 2018.

[4] Y. Li, Y. Shang, and Y. Yang, "Clustering coefficients of large networks," *Information Sciences*, vol. 382, pp. 350–358, 2017.

[5] Q. Liu, H. Li, X. Liu, and M. Jiang, "Information networks in the stock market based on the distance of the multi-attribute dimensions between listed companies," *Physica A: Statistical Mechanics and its Applications*, vol. 496, pp. 505–513, 2018.

[6] M. Riondato and E. Upfal, "Abra: Approximating betweenness centrality in static and dynamic graphs with rademacher averages," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 5, p. 61, 2018.

[7] 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.

[8] 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. 14 151–14 158, 2018.

[9] M. Bressan, F. Chierichetti, R. Kumar, S. Leucci, and A. Panconesi, "Motif counting beyond five nodes," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 4, p. 48, 2018.

[10] W.-J. Xie, R.-Q. Han, and W.-X. Zhou, "Tetradic motif profiles of horizontal visibility graphs," *Communications in Nonlinear Science and Numerical Simulation*, 2019.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

[11] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.

[12] L. Romijn, B. Ó. Nualláin, and L. Torenvliet, "Discovering motifs in real-world social networks," in *International Conference on Current Trends in Theory and Practice of Informatics*. Springer, 2015, pp. 463–474.

[13] M. B. Z. Joveini and J. Sadri, "Application of fractal theory on motifs counting in biological networks," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 2, pp. 613–623, 2018.

[14] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón *et al.*, "Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond," *Nucleic Acids Research*, vol. 44, no. D1, pp. D133–D143, 2015.

[15] A. Verfaillie, H. Imrichová, B. Van de Sande, L. Standaert, V. Christiaens, G. Hulselmans, K. Herten, M. N. Sanchez, D. Potier, D. Svetlichnyy *et al.*, "iRegulon: from a gene list to a gene regulatory network using large motif and track collections," *PLoS Computational Biology*, vol. 10, no. 7, p. e1003731, 2014.

[16] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, p. 101, 2004.

[17] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.

[18] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, "Intensity and coherence of motifs in weighted complex networks," *Physical Review E*, vol. 71, no. 6, p. 065103, 2005.

[19] P. Ribeiro, F. Silva, and M. Kaiser, "Strategies for network motifs discovery," in *Fifth IEEE International Conference on E-Science*. IEEE, 2009, pp. 80–87.

[20] N. T. L. Tran and C.-H. Huang, "A survey of motif finding web tools for detecting binding site motifs in chip-seq data," *Biology Direct*, vol. 9, no. 1, p. 4, 2014.

[21] G. Ciriello and C. Guerra, "A review on models and algorithms for motif discovery in protein–protein interaction networks," *Briefings in Functional Genomics and Proteomics*, vol. 7, no. 2, pp. 147–156, 2008.

[22] M. K. Das and H.-K. Dai, "A survey of DNA motif finding algorithms," in *BMC Bioinformatics*, vol. 8, no. 7. BioMed Central, 2007, p. S21.

[23] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, "On the uniform generation of random graphs with prescribed degree sequences," *Quantitative Biology*, 2004.

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

[25] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield, "Efficient graphlet counting for large networks," in *IEEE International Conference on Data Mining (ICDM)*. IEEE, 2015, pp. 1–10.

[26] M. E. Silva, P. Paredes, and P. Ribeiro, "Network motifs detection using random networks with prescribed subgraph frequencies," in *Workshop on Complex Networks CompleNet*. Springer, 2017, pp. 17–29.

[27] F. Picard, J.-J. Daudin, M. Koskas, S. Schbath, and S. Robin, "Assessing the exceptionality of network motifs," *Journal of Computational Biology*, vol. 15, no. 1, pp. 1–20, 2008.

[28] E. Ziv, R. Koytcheff, M. Middendorf, and C. Wiggins, "Systematic identification of statistically significant network measures," *Physical Review E*, vol. 71, no. 1, p. 016110, 2005.

[29] J. Berg and M. Lässig, "Local graph alignment and motif search in biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 41, pp. 14 689–14 694, 2004.

[30] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Topological generalizations of network motifs," *Physical Review E*, vol. 70, no. 3, p. 031909, 2004.

[31] J.-R. Kim, Y. Yoon, and K.-H. Cho, "Coupled feedback loops form dynamic motifs of cellular networks," *Biophysical Journal*, vol. 94, no. 2, pp. 359–365, 2008.

[32] P. Bloem and S. de Rooij, "Large-scale network motif learning with compression," *CoRR, vol. abs/1701.02026*, 2017.

[33] L. Parida, "Discovering topological motifs using a compact notation," *Journal of Computational Biology*, vol. 14, no. 3, pp. 300–323, 2007.

[34] J. Huan, W. Wang, J. Prins, and J. Yang, "Spin: mining maximal frequent subgraphs from graph databases," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2004, pp. 581–586.

[35] S. Fu-li, L. Yong-lin, and Z. Yi-fan, "A military communication super-network structure model for netcentric environment," in *International Conference on Computational and Information Sciences (ICCIS)*. IEEE, 2010, pp. 33–36.

[36] F. Shi, C. Li, D. Qin, Y. Zhu, and F. Yang, "A complexity measure for military communication networks," in *Proceedings of Military Communications Conference(MILCOM)*. IEEE, 2011, pp. 1708–1713.

[37] H. Han, W. Liu, and L. Wu, "The measurement of complex network based on motif," *Acta Physica Sinica*, vol. 62, no. 168904, 2013.

[38] L. Egghe and L. Leydesdorff, "The relation between pearson's correlation coefficient r and salton's cosine measure," *Journal of the Association for Information Science and Technology*, vol. 60, no. 5, pp. 1027–1036, 2009.

[39] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[40] J. C. Claussen, "Offdiagonal complexity: A computationally quick complexity measure for graphs and networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 375, no. 1, pp. 365–373, 2007.

[41] T. Wilhelm and J. Hollunder, "Information theoretic description of networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 385, no. 1, pp. 385–396, 2007.

[42] J. Kim and T. Wilhelm, "What is a complex graph?" *Physica A: Statistical Mechanics and Its Applications*, vol. 387, no. 11, pp. 2637–2652, 2008.

[43] M. Dehmer, N. Barbarini, K. Varmuza, and A. Graber, "A large scale analysis of information-theoretic network complexity measures using chemical structures," *PLoS One*, vol. 4, no. 12, p. e8057, 2009.

[44] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.

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

[46] M. Kuramochi and G. Karypis, "Finding frequent patterns in a large sparse graph," *Data mining and knowledge discovery*, vol. 11, no. 3, pp. 243–271, 2005.

[47] V. Q. Marinho, G. Hirst, and D. R. Amancio, "Authorship attribution via network motifs identification," in *5th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, 2016, pp. 355–360.

[48] J. Liu, H. A. Abbass, D. G. Green, and W. Zhong, "Motif difficulty (md): a predictive measure of problem difficulty for evolutionary algorithms using network motifs," *Evolutionary Computation*, vol. 20, no. 3, pp. 321–347, 2012.

[49] 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 Bioinformatics*, vol. 10, no. 1, p. 318, 2009.

[50] J. A. Grochow and M. Kellis, "Network motif discovery using subgraph enumeration and symmetry-breaking," in *Annual International Conference on Research in Computational Molecular Biology*. Springer, 2007, pp. 92–106.

[51] C. Schmidt, T. Weiss, C. Komusiewicz, H. Witte, and L. Leistritz, "An analytical approach to network motif detection in samples of networks with pairwise different vertex labels," *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012.

[52] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 601–610.

[53] G. Han and H. Sethu, "Waddling random walk: Fast and accurate mining of motif statistics in large graphs," in *IEEE International Conference on Data Mining*, 2017.

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

[55] E. Wong, B. Baur, S. Quader, and C.-H. Huang, "Biological network motif detection: principles and practice," *Briefings in Bioinformatics*, vol. 13, no. 2, pp. 202–215, 2011.

[56] G. Wu, M. Harrigan, and P. Cunningham, "Characterizing wikipedia pages using edit network motif profiles," in *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*. ACM, 2011, pp. 45–52.

[57] R. Guimera, M. Sales-Pardo, and L. A. Amaral, "Classes of complex networks defined by role-to-role connectivity profiles," *Nature Physics*, vol. 3, no. 1, p. 63, 2007.

[58] K. Juszczyszyn, P. Kazienko, and K. Musiał, "Local topology of social network based on motif analysis," in *International Conference on Knowledge-*

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

Based and Intelligent Information and Engineering Systems. Springer, 2008, pp. 97–105.

[59] K. Juszczyszyn, K. Musiał, P. Kazienko, and B. Gabrys, "Temporal changes in local topology of an email-based social network," *Computing and Informatics*, vol. 28, no. 6, pp. 763–779, 2012.

[60] X. Xu, J. Zhang, and M. Small, "Superfamily phenomena and motifs of networks induced from time series," *Proceedings of the National Academy of Sciences*, vol. 105, no. 50, pp. 19 601–19 605, 2008.

[61] N. Przulj, "Biological network comparison using graphlet degree distribution." *Bioinformatics*, vol. 23, no. 2, p. e177, 2007.

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

[63] C. E. Tsourakakis, J. Pachocki, and M. Mitzenmacher, "Scalable motif-aware graph clustering," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 1451–1460.

[64] P. Li, H. Dau, G. Puleo, and O. Milenkovic, "Motif clustering and overlapping clustering for social network analysis," in *IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2017, pp. 1–9.

[65] H. Yin, A. R. Benson, and J. Leskovec, "Higher-order clustering in networks," *Physical Review E*, vol. 97, no. 5, p. 052306, 2018.

[66] C. Pizzuti and A. Socievole, "Multiple network motif clustering with genetic algorithms," in *Italian Workshop on Artificial Life and Evolutionary Computation*. Springer, 2017, pp. 296–307.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2926752, IEEE Access

FENG XIA (M'07-SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He was a Research Fellow with the Queensland University of Technology, Australia. He is currently a Full Professor with the School of Software, Dalian University of Technology, China. He has published two books and over 200 scientific papers in international journals and conferences. His research interests include data science, big data, knowledge management, network science, and mobile social networks. He is a Senior Member of the IEEE and ACM, and a member of AAAS.

HAORAN WEI received the B.S. degree in software engineering from the Hefei University of Technology, China, in 2017. She is currently pursuing the master's degree with The Alpha Lab, School of Software the School of Software, Dalian University of Technology, China. Her research interests include scholarly big data and network science.

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

DA ZHANG received the B.E. degree in Software Engineering (2010) from Dalian University of Technology with First-Class Honors, and the M.S. degree in Computer Science and Engineering (2012) from the Ohio State University, U.S.A. She has been pursuing her Ph.D. in Electrical and Computer Engineering at the University of Miami (UM), Coral Gables, Florida since 2015. She is currently a Ph.D. candidate and a teaching assistant at UM. Her research interests include Knowledge Graph Management, Big Data, Web Semantics and Deep Learning.

BO XU received the BSc and PhD degrees from the Dalian University of Technology, China, in 2007 and 2014, respectively. She is currently a lecture in School of Software at the Dalian University of Technology. Her current research interests include social computing, machine learning, literature data mining, and natural language processing.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.