Multiple Network Motif Clustering with Genetic Algorithms

Clara Pizzuti(✉) and Annalisa Socievole

National Research Council of Italy (CNR), Institute for High Performance Computing and Networking (ICAR), Via Pietro Bucci, 7-11C, 87036 Rende (CS), Italy
{clara.pizzuti,annalisa.socievole}@icar.cnr.it

Abstract. The definition of community, usually, relies on the concept of edge density. Network motifs, however, have been recognized as fundamental building blocks of networks and, similarly to edges, may give insights for uncovering communities in complex networks. In this work, we propose a novel approach for identifying communities of network motifs. Differently from previous approaches, our method focuses on searching communities where nodes simultaneously participate in several types of motifs. Based on a genetic algorithm, the method finds a number of communities by minimizing the concept of multiple-motifs conductance. Simulations on a real-world network show that the proposed algorithm is able to better capture the real modular structure of the network, outperforming both motifs-based and classic community detection algorithms.

Keywords: Community detection · Network motifs
Evolutionary techniques · Genetic algorithm

1 Introduction

Complex networks contain small subgraphs named *network motifs* [11], which are pattern of interconnections recurring more frequently than expected in a random network. The frequency of a motif describes the number of times this motif appears within the network. High frequencies of certain motifs are possible due to the important functions they play in a network. For example, the *feed forward loop* and the *bifan* motifs shown in Fig. 1(a) and (d), respectively, have been found to be highly frequent into the genetic regulation networks of *E. coli* and *S. Cerevisiae*, as well as into the *C. elegans* neurons network. It is worth noting that multiple motifs usually coexist within a network. Figure 2 shows a subgraph of *Florida Bay food web* network [18], where different microorganisms interact through multiple motifs. We highlight here three types of network motifs: M_5, M_6 (Fig. 1(b)) and M_8 (Fig. 1(c)). In motif M_5, Water POC serves as energy source for Free Bacteria and Meroplankton, and Meroplankton for Free Bacteria. In motif M_6, Free Bacteria acts also as energy source for Arcatia tonsa, and both nodes are served by Input. Finally, interaction patterns like Input serving Water flagellates and Water ciliates (motif M_8) occur many times.

M. Pelillo et al. (Eds.): WIVACE 2017, CCIS 830, pp. 296–307, 2018.
https://doi.org/10.1007/978-3-319-78658-2_22

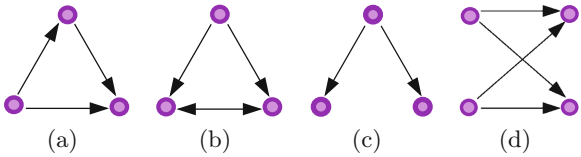

Fig. 1. (a) M_5 (feed-forward loop), (b) M_6, (c) M_8, and (d) M_{bifan} motifs.

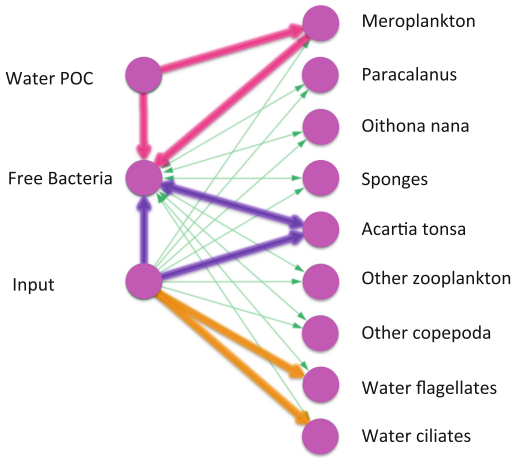

Fig. 2. Multiple motifs coexisting in a subgraph of Florida Bay food web network: M_5, M_6 and M_8. The edges composing these three motifs are highlighted in pink, purple and orange, respectively. (Color figure online)

Although network motifs have been recognized as "fundamental units of networks" [3], few studies explore the role these subpatterns have in community detection. Arenas et al. [1] show how motifs can be used to define a *motif-based modularity*, i.e. how motif-based modules present more motifs than a random division. Specifically, they extended the original definition of modularity introduced by Girvan and Newman [6] to deal with classes of motifs, and showed that the detected partitions are different with respect to those obtained by optimizing the classical modularity. A spectral method based on the generalized modularity [17] has been proposed by the same authors, and the differences between the obtained community structures on several networks are highlighted. In a recent work, Benson et al. [2] proposed a tensor spectral clustering method that clusters nodes according to the motif specified in input by the user. First, the higher-order structures involving multiple nodes are encoded by means of tensors (i.e., multidimensional matrices). Then, the method searches a partitioning that does not cut the motifs. Another work [3] by the same authors, described in detail in

the next section, extends the concept of *conductance* [16] to network motifs for finding cluster of motifs with low *motif conductance*.

One of the main drawbacks of the aforementioned approaches is that the number of communities must be fixed in advance. In a previously work [14], we proposed *MotifGA*, an evolutionary motifs-based algorithm for community detection using *Genetic Algorithms (GAs)* [7] and a type of motif as input for discovering a number of motif-based communities minimizing motif conductance. Here and in all the previous cited related works, motif-based clustering is thus performed fixing a type of motif and exploring the communities based only on that motif, without considering the coexistence of multiple motifs.

In this paper, we propose *M-MotifGA*, a genetic algorithm for detecting communities in complex networks simultaneously considering different motifs. The method evolves a population of individuals by minimizing the concept of *multiple-motifs conductance*, and finds a partition of the network into k communities, with k determined by the best local solution optimizing the multiple-motifs conductance as fitness function. A comparison with the approach of Benson et al. [3], with a variant of this approach we developed here for taking into account multiple motifs, and with the two best known community detection methods *Louvain* [4] and *Infomap* [15] shows that *M-MotifGA* obtains results better than those found by the other state-of-the-art methods.

The paper is organized as follows. Section 2 introduces the concepts of conductance when motifs are considered and defines the problem we tackle. Section 3 describes our method. Section 4 details the dataset used to perform our experiments and the results obtained. Finally, Sect. 5 concludes the paper.

2 Network Motif Clustering

In this section, we start recalling the concepts of *network motif*, *conductance*, *motif conductance* and *multiple-motifs conductance*. Then, we describe the method proposed by Benson et al. [3] and the introduction of multiple motifs within their method.

Given a graph $G = (V, E)$ with weights W, $n =\mid V \mid$ number of vertices, and $m =\mid E \mid$ number of edges, a *motif* M of G on r nodes $\{v_1, \ldots, v_r\}$, represented by a sub-adjacency matrix of size $r \times r$, is defined as a subgraph of G presenting a particular pattern of interconnections. Figure 1 shows three types of motifs among three nodes (Fig. 1(a), (b), and (c)) and a motif involving four nodes (Fig. 1(d)). Their labeling follows the same convention adopted in [3].

Given the diagonal degree matrix D of G defined as $D_{ii} = \sum_{j=1}^{n} W_{ij}$, and a set $S \subset V$ of nodes, the *cut* of S, denoted $cut(S)$, is defined as the sum of edge weights having one endpoint in S and the other in $\overline{S} = V - S$:

$$cut(S) = \sum_{i \in S, j \in \overline{S}} W_{ij} \tag{1}$$

The *conductance* of S is defined as

$$\phi(S) = \frac{cut(S)}{min(vol(S), vol(\overline{S}))} \tag{2}$$

where $vol(S) = \sum_{i \in S} D_{ii}$ is the weighted sum of edge end points in S.

By substituting an edge with a motif instance of M, the conductance of S can be generalized to motifs as follows

$$\phi_M(S) = \frac{cut_M(S)}{min(vol_M(S), vol_M(\overline{S}))} \tag{3}$$

where $\phi_M(S)$ is defined *motif conductance*, $cut_M(S)$ is the number of motif instances of M with at least a node in S and another in $\overline{S}$, and $vol_M(S)$ is the number of instances of M contained in S.

Problem definition (single-motif). Fixed a network motif M, find a set of nodes S such that: (1) they participate in as many instances of M as possible, and (2) cutting instances of M is avoided, i.e. all the nodes of M should belong to either S or $\overline{S}$.

Benson et al. [3] proposed a method for partitioning V into the complementary sets S and $\overline{S}$ that, given a motif M, minimizes the motif conductance $\phi_M(S)$. The method works on the *motif adjacency matrix* W_M, where each element represents the number of times two nodes appear in an instance of M. When there are nodes that do not participate in any motif, these nodes are discarded from W_M. Then, the eigenvector corresponding to the second smallest eigenvalue of the normalized motif Laplacian matrix is computed. The components of the eigenvector generate an ordering of nodes, which produces nested sets of nodes. The set of nodes with the smallest motif conductance is proven to be a near-optimal partition. Further details on the approach can be found in [3]. For obtaining a partition with more than two communities, the method, named *Motif Recursive bi-partitioning* (*MRbi-part*), can be recursively executed on S and $\overline{S}$, until the desired number of clusters is obtained.

When considering M_1, M_2, ..., M_q motifs simultaneously, the *multiple-motifs conductance* is defined as

$$\phi_{MM}(S) = \frac{\sum_{j=1}^{q} \alpha_j cut_{M_j}(S)}{min(\sum_{j=1}^{q} \alpha_j vol_{M_j}(S), \sum_{j=1}^{q} \alpha_j vol_{M_j}(\overline{S}))} \tag{4}$$

where each $\alpha_j \geq 0$ gives a weight to the impact of motif M_j on the considered network.

Problem definition (multiple-motifs). Given a set of q network motifs M_1, M_2, ..., M_q, find a set of nodes S such that (1) they simultaneously participate in as many instances of all the considered motifs as possible, and (2) cutting instances of any M_j, $j = 1, ..., q$ are avoided.

In the next section, we propose to solve the problem of finding a division on a network based on multiple motifs by applying a Genetic Algorithm. Specifically,

the proposed algorithm minimizes the multiple-motifs conductance computed on the motif adjacency matrices of the single motifs, associated with the graph G representing the network. For comparing our method to another method based on multiple motifs, we also modified *MRbi-part* extending its code such that the multiple-motifs conductance is the measure to minimize. As such, we considered a *weighted motif adjacency matrix* $W_{Mw} = \sum_{j=1}^{q} \alpha_j W_{M_j}$ for running the method. We denominate this extension as $MRbi\text{-}part_{MM}$. It is worth noting that, differently from the methods by Benson et al., our method does not need a prior setting of the number of communities to find. This number is automatically provided by decoding the solution obtained by the method, i.e. the solution with the lowest local optimum value of conductance.

3 *M-MotifGA* Description

The algorithm we propose, named *M-MotifGA* is based on our previous work [14], where we proposed *MotifGA*, an approach to motif network clustering exploiting a genetic algorithm, that evolves a population of individuals by minimizing motif conductance. Similarly to *MotifGA*, *M-MotifGA* obtains the simultaneous partition of a network into k communities, with k determined by the best local solution optimizing the fitness function. However, differently from *MotifGA*, the fitness function used in *M-MotifGA* is the multiple-motifs conductance.

A GA-based method basically evolves a population of individuals initialized at random, and performs variation and selection operators to increase the value of a criterion function, while exploring the search space during the optimization process. *M-MotifGA* uses the locus-based adjacency representation [13] for representing the problem, uniform crossover and neighbor-based mutation for evolving individuals. In the locus-based representation, an individual I is represented as a vector of n genes. Each gene can assume a value j in the range $\{1, \ldots, n\}$: when a value j is assigned to the ith node, nodes i and j are linked. A decoding step identifies all the connected components of the graph which correspond to the network division in communities. Uniform crossover generates a random binary mask of length equal to the number of nodes, and an offspring is obtained by selecting from the first parent the genes in the mask set to 0, and from the second parent the genes in the mask set to 1. Finally, the mutation operator randomly changes the value j of a gene to one of its neighbors.

M-MotifGA receives in input the graph $G = (V, E)$ and the set of motifs M_1, M_2, ..., M_q, and performs the following steps:

1. compute the motif adjacency matrices W_{M_1}, W_{M_2}, ..., W_{M_q};
2. take the largest connected component $W_{M_j}^{max}$ of W_{M_j} for each motif M_j of the q motifs;
3. obtain the weighted graph $G_{M_j} = (V_{M_j}, E_{M_j})$ corresponding to $W_{M_j}^{max}$ for each $1 \leq j \leq q$;
4. compute the weighted graph $G_M = \sum_{j=1}^{q} G_{M_j}$;

5. run the GA on G_M for a number of iterations by using multiple-motifs conductance as fitness function to minimize, uniform crossover and neighbor mutation as variation operators;
6. obtain the partition $C = \{C_1, \ldots, C_k\}$ corresponding to the solution with the lowest fitness value;
7. merge two communities if the number of inter-cluster connections is higher than the number of intra-cluster connections.

Note that the weighted graphs G_{M_j} associated with the largest connected components of the motif adjacency matrices may have different numbers of nodes. In this case, before Step 4, the algorithm computes the subset of nodes which are common to all the G_{M_j} graphs. Then, the matrices G_{M_j} will be reorganized in order to contain only the rows and the columns related to that subset of nodes.

In the next section we present the results obtained by our algorithm and compare them with those returned by state-of-the-art methods. Moreover, we also investigate a variant of our approach, named *MS-MotifGA*, that uses as fitness function the sum of the motif conductance of the single motifs, that is $\phi_{MS}(S) = \phi_{M_1}(S) + \phi_{M_2}(S) + \ldots + \phi_{M_q}(S)$.

4 Experimental Evaluation

To validate our algorithm, we performed several simulations using Matlab 2015b and the Global Optimization Toolbox. Specifically, we compared our algorithm with other well-known state-of-the-art algorithms in terms of NMI [5], ARI [9] and $F1$ [10] indexes. The results for *M-MotifGA* have been averaged over 10 runs of the algorithm, setting the population size to 100, the number of generations to 200, the mutation rate to 0.2, and the crossover rate to 0.8. These parameter values have been fixed by employing a trial-and-error procedure on the benchmark data set. Moreover, for computing the multiple-motifs conductance, we equally weighted all motifs using $\alpha_j = 1$. For *MRbi-part* we set to 4 the number of communities to find, as suggested by Benson et al. for this dataset, since a higher number of communities would give higher motif conductance values and, thus, worse results for their algorithm. Specifically, we applied the motif recursive bipartitioning method twice in order to obtain the desired number of communities. The following subsections describe the dataset and performance indexes used, and the algorithms taken into account for testing the effectiveness of *M-MotifGA*.

4.1 Dataset

We analyze the **Florida Bay food web** dataset containing the data of an ecosystem food web. Converting these data into a network graph, nodes can be considered organisms and species, and edges the directed carbon exchange between species. For clustering this network, we consider the motifs M_5, M_6 and M_8 shown in Fig. 1. $M5$, considered a building block for food webs, describes the

hierarchical flow of energy between species i and j which are energy sources for species k, while i is an energy source for both. M_6, on the contrary, models two species that exchange energy and compete to receive energy from a third specie. This motif has been shown to be prevalent within this network, resulting in a rich high-order modular structure. Finally, M_8 corresponds to a single specie feeding two non-interacting species.

The original Florida Bay food web network is composed by 128 nodes and 2106 edges. For detecting communities, we consider a subset of 62 nodes for which the ground truths are known. Specifically, two ground truths, denoted as $GT1$ and $GT2$, are available and are relative to the two large connected components resulting from the analysis of the adjacency matrix of motif M_6. Table 1 shows the 50 nodes corresponding to the first component and the 12 nodes of the second component. The remaining 66 nodes are isolated. In $GT1$ nodes are classified into 11 different categories (*'demersal producer'*, *'seagrass producer'*, *'algae producer'*, *'microbial microfauna'*, *'zooplankton microfauna'*, *'sediment organism microfauna'*, *'macroinvertebrates'*, *'pelagic fishes'*, *'benthic fishes'*, *'demersal fishes'*, and *'detritus'*). In $GT2$, on the contrary, nodes are categorized into 7 groups: *'producer'*, *'microfauna'*, *'macroinvertebrates'*, *'pelagic fishes'*, *'benthic fishes'*, *'demersal fishes'*, and *'detritus'*. Basically, $GT2$ considers all the *producer* and *microfauna* subcategories of $GT1$ as unique macro categories.

The largest connected components of the adjacency matrices for motifs M_5 and M_8 have 127 and 128 nodes, respectively. Since both motif adjacency matrices contain the 62 nodes for which the ground truths are known, we consider only the sub-matrices corresponding to this set of 62 nodes when dealing with motifs $M5$ and $M8$.

4.2 Performance Indexes

To assess the quality of the solutions, we use the following evaluation measures, well known in the literature:

NMI. The normalized mutual information $NMI(A, B)$ [5] of two divisions A and B of a network is defined as follows. Let C be the confusion matrix whose element C_{ij} is the number of nodes of community i of the partition A that are also in the community j of partition B.

$$NMI(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} C_{ij} log(C_{ij}n/C_{i.}C_{.j})}{\sum_{i=1}^{c_A} C_{i.} log(C_{i.}/n) + \sum_{j=1}^{c_B} C_{.j} log(C_{.j}/n)} \tag{5}$$

where c_A (c_B) is the number of groups in partition A (B), $C_{i.}$ ($C_{.j}$) is the sum of the elements of C in row i (column j), and n is the number of nodes. If $A = B$, $NMI(A, B) = 1$. If A and B are completely different, $NMI(A, B) = 0$.

F1 score. This measure [10] is calculated by using the Precision (P) and Recall (R) measures as $F1 = \frac{2RP}{R+P}$, where $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$. True Positive (TP) refers to the number of nodes which are correctly assigned to communities,

Table 1. Florida Bay food web ground truths (GT1 and GT2) for the two large connected components of the motif $M6$ adjacency matrix.

Node ID	Species	Component	GT1	GT2
8	'Benthic Phytoplankton'	1	Demersal Producer	Producer
9	'Thalassia'	1	Seagrass Producer	Producer
10	'Halodule'	1	Seagrass Producer	Producer
11	''Syringodium'	1	Seagrass Producer	Producer
13	'Drift Algae'	1	Algae Producer	Producer
14	'Epiphytes'	1	Algae Producer	Producer
24	'Benthic Flagellates'	1	Sediment Organism Microfauna	Microfauna
25	'Benthic Ciliates'	1	Sediment Organism Microfauna	Microfauna
26	'Meiofauna'	1	Sediment Organism Microfauna	Microfauna
29	'Other Cnidaridae'	1	Macroinvertebrates	Macroinvertebrates
30	'Echinoderma'	1	Macroinvertebrates	Macroinvertebrates
31	'Bivalves'	1	Macroinvertebrates	Macroinvertebrates
32	'Detritivorous Gastropods'	1	Macroinvertebrates	Macroinvertebrates
34	'Predatory Gastropods'	1	Macroinvertebrates	Macroinvertebrates
35	'Detritivorous Polychaetes'	1	Macroinvertebrates	Macroinvertebrates
36	'Predatory Polychaetes'	1	Macroinvertebrates	Macroinvertebrates
37	'Suspension Feeding Polych'	1	Macroinvertebrates	Macroinvertebrates
38	'Macrobenthos'	1	Macroinvertebrates	Macroinvertebrates
39	'Benthic Crustaceans'	1	Macroinvertebrates	Macroinvertebrates
40	'Detritivorous Amphipods'	1	Macroinvertebrates	Macroinvertebrates
41	'Herbivorous Amphipods'	1	Macroinvertebrates	Macroinvertebrates
42	'Isopods'	1	Macroinvertebrates	Macroinvertebrates
43	'Herbivorous Shrimp'	1	Macroinvertebrates	Macroinvertebrates
44	'Predatory Shrimp'	1	Macroinvertebrates	Macroinvertebrates
45	'Pink Shrimp'	1	Macroinvertebrates	Macroinvertebrates
48	'Detritivorous Crabs'	1	Macroinvertebrates	Macroinvertebrates
49	'Omnivorous Crabs'	1	Macroinvertebrates	Macroinvertebrates
50	'Predatory Crabs'	1	Macroinvertebrates	Macroinvertebrates
51	'Callinectus sapidus'	1	Macroinvertebrates	Macroinvertebrates
57	'Sardines'	1	Pelagic Fishes	Pelagic Fishes
58	'Anchovy'	1	Pelagic Fishes	Pelagic Fishes
59	'Bay Anchovy'	1	Pelagic Fishes	Pelagic Fishes
60	'Lizardfish'	1	Benthic Fishes	Benthic Fishes
61	'Catfish'	1	Benthic Fishes	Benthic Fishes
62	'Eels'	1	Demersal Fishes	Demersal Fishes
63	'Toadfish'	1	Benthic Fishes	Benthic Fishes
64	'Brotalus'	1	Demersal Fishes	Demersal Fishes
65	'Halfbeaks'	1	Pelagic Fishes	Pelagic Fishes
66	'Needlefish'	1	Pelagic Fishes	Pelagic Fishes
68	'Goldspotted killifish'	1	Demersal Fishes	Demersal Fishes
69	'Rainwater killifish'	1	Demersal Fishes	Demersal Fishes
72	'Silverside'	1	Pelagic Fishes	Pelagic Fishes
91	'Mullet'	1	Pelagic Fishes	Pelagic Fishes
93	'Blennies'	1	Benthic Fishes	Benthic Fishes
94	'Code Goby'	1	Benthic Fishes	Benthic Fishes
95	'Clown Goby'	1	Benthic Fishes	Benthic Fishes
96	'Flatfish'	1	Benthic Fishes	Benthic Fishes
99	'Other Pelagic Fishes'	1	Pelagic Fishes	Pelagic Fishes

(*continued*)

Table 1. (*continued*)

Node ID	Species	Component	GT1	GT2
100	'Omnivorous Ducks'	1	Demersal Fishes	Demersal Fishes
124	'Benthic POC'	1	Detritus	Detritus
15	'Free Bacteria'	2	Microbial Microfauna	Microfauna
16	'Water Flagellates'	2	Microbial Microfauna	Microfauna
17	'Water Cilitaes'	2	Microbial Microfauna	Microfauna
18	'Acartia Tonsa'	2	Zooplankton Microfauna	Microfauna
19	'Oithona nana'	2	Zooplankton Microfauna	Microfauna
20	'Paracalanus'	2	Zooplankton Microfauna	Microfauna
21	'Other Copepoda'	2	Zooplankton Microfauna	Microfauna
22	'Meroplankton'	2	Zooplankton Microfauna	Microfauna
23	'Other Zooplankton'	2	Zooplankton Microfauna	Microfauna
27	'Sponges'	2	Macroinvertebrates	Macroinvertebrates
123	'Water POC'	2	Detritus	Detritus
126	'Input'	2	Detritus	Detritus

False Positive (FP) refers to the nodes which are incorrectly assigned to communities, and False Negatives (FN) refers to the set of nodes which are incorrectly not assigned to the proper communities. F1 value reaches its best value at 1 and worst at 0.

Adjusted Rand Index. The Adjusted Rand Index (ARI) is a normalized version of the *Rand Index (RI)*[9] which simply assesses the degree of agreement between two partitions A and B. Let n_{11} be the number of pairs appearing in the same cluster in both A and B, n_{00} the number of pairs that appear in different clusters in both A and B, n_{10} the number of pairs appearing in the same cluster in A but in different clusters in B, and n_{01} the number of pairs that are in the same cluster in B and not in A. Then

$$ARI(A,B) = \frac{2(n_{00}n_{11} - n_{01}n_{10})}{(n_{00}+n_{01})(n_{01}+n_{11}) + (n_{00}+n_{10})(n_{10}+n_{11})} \tag{6}$$

4.3 Algorithms for Community Detection

We compare the two strategies of *M-MotifGA*, namely *MM-MotifGA*, in which the fitness function used is $\phi_{MM}(S)$, and *MS-MotifGA*, in which the fitness function is the sum of the single motif conductances, with the motifs-based *MRbi-part*, both in the case in which this last algorithm uses a single motif to detect communities and in the case of multiple motifs jointly used. We also compare *M-MotifGA* to two benchmark algorithms not using motifs: Louvain [4] and Infomap [15]. Louvain basically tries to optimize the modularity [12] of a partition through a greedy optimization technique. First, small communities are searched by optimizing modularity locally. Then, a new network whose nodes are the communities are built and these steps are repeated until a hierarchy of high-modularity communities is obtained. Infomap, on the contrary, exploits the principles of information theory characterizing the problem of community

detection as the problem of finding a description of minimum information of a random walk on the graph. Maximizing the Minimum Description Length [8] objective function, Infomap quickly provides an approximation of the optimal solution.

4.4 Results

Table 2(a)–(b) shows the NMI, ARI and $F1$ values obtained for the two ground truths of the Florida Bay food web results. The statistical significance of the results has been been checked by performing a t-test at the 5% significance level. The test rejected the null hypothesis that the values come from populations with equal means, and returned p-values, on average, below 0.1E-5.

For *MM-MotifGA* and *MS-MotifGA* we report both the average value and the standard deviation (in parenthesis) of the evaluation measures. For *MS-MotifGA*, we investigated as fitness function $\phi_{MS}(S) = \phi_{M_5}(S)+\phi_{M_6}(S)+\phi_{M_8}(S)$. On the ground truth $GT1$, *MM-MotifGA* outperforms all the other community detection schemes finding a number of communities ranging from 7 to 10. Similarly, *MS-MotifGA*, finds solutions with 7, 8, 9 and 10 communities, outperforming all the other methods. Considering *MS-MotifGA*, however, we observe that the strategy to sum the single motif conductances as fitness function to optimize, results in NMI, ARI and $F1$ values lower than *MM-MotifGA*. As such, we conclude that if we explore separately the three motifs and then we recombine them by summing their conductances to obtain the function to optimize, the algorithm does not take into account the intersection which could exist between motifs in terms of edges. This intersection, as in the case of motifs M_5 and M_8, and M_6 and M_8 for example, considered when jointly analyzing multiple motifs, is able to

Table 2. Florida Bay food web results.

		MM-MotifGA	***MS-MotifGA***	***MRbi-part***$_{M_5}$	***MRbi-part***$_{M_6}$
$GT1$	NMI	0.9241 (0.0756)	0.8602 (0.0781)	0.4392	0.504
	ARI	0.8451 (0.1923)	0.6879 (0.2141)	0.1388	0.3005
	$F1$	0.8765 (0.1489)	0.754 (0.1646)	0.3149	0.4437
$GT2$	NMI	0.8367 (0.1127)	0.6844 (0.1054)	0.3214	0.4822
	ARI	0.7039 (0.1798)	0.3886 (0.1106)	0.1045	0.3265
	$F1$	0.7756 (0.1329)	0.5549 (0.0727)	0.3087	0.4802

(a)

		MRbi-part$_{M_8}$	***MRbi-part***$_{MM}$	**Louvain**	**Infomap**
$GT1$	NMI	0.4197	0.3406	0.3879	0.4035
	ARI	0.1203	0.1291	0.2207	0.1423
	$F1$	0.2949	0.2962	0.4068	0.31
$GT2$	NMI	0.3573	0.2829	0.3034	0.3471
	ARI	0.1332	0.1241	0.2229	0.1592
	$F1$	0.3244	0.3101	0.434	0.3416

(b)

provide more meaningful communities, as the results show. Analyzing all the algorithms by Benson et al. where the number of communities has been set to 4, the communities found result in significantly lover values of the evaluation measures we considered. It is worth noting that considering only M_5, M_6 or M_8 for clustering nodes does not produce satisfying results compared to our multiple-motifs strategies. Moreover, when jointly considering all the motifs as in *MRbi-part*$_{MM}$, the algorithm performs even worse than the single-motif strategies *MRbi-part*$_{M_5}$, *MRbi-part*$_{M_6}$, and *MRbi-part*$_{M_8}$. As such, we conclude that when the number of communities needs to be fixed in input as for *MRbi-part*, detecting clusters of multiple motifs using multiple-motifs conductance as the function to be minimized may lead to suboptimal results. Finally, comparing our method to *Louvain* and $Infomap$, which do not exploit motifs, we observe that also these methods are not able to find a good match with the ground truth. Focusing on the largest community of the ground truth (i.e., the *macroinvertebrates*) including 21 nodes, for example, we observe that *MM-MotifGA* perfectly matches it in all the runs of the algorithm. *Louvain* distributes the nodes into 4 different communities: 2 groups with 7 nodes are inserted into different communities, 3 nodes into another one, and the remaining nodes into another community. Finally, $Infomap$ inserts all the 21 nodes into a unique but larger community including other nodes.

On the ground truth $GT2$, we obtain similar results. *MM-MotifGA* still outperforms all other methods, resulting in the highest NMI, ARI and $F1$ values finding solutions with 5 or 6 communities. Overall, for all the algorithms, we observe NMI, ARI and $F1$ values for $GT2$ are lower than the values obtained for $GT1$. This behavior was also observed in our previous work [14] and it is probably due to the merging of some specie categories done on $GT2$ to create macro-categories which do not perfectly reflect the modular structure of the network.

5 Conclusion

In this paper, we have proposed *M-MotifGA*, a method for discovering communities composed by multiple motifs. Based on a genetic algorithm, our method simultaneously considers different motifs for searching a partition with a number of communities minimizing the multiple-motifs conductance as fitness function. Simulations on the Florida bay food web network show that *M-MotifGA* results in NMI, $F1$ and ARI values much higher than both the single-motif and the multiple-motif based analyzed strategies, Louvain and Infomap. Specifically, we have observed that for better matching the underlying real communities, not only multiple motifs should be simultaneously considered, but also fixing the number of communities to obtain as in Benson et al. [3] does not fully exploit the benefits of considering multiple motifs. As future work, we plan to extend our experiments to other datasets to further validate our method. We also intend to explore how community detection can be performed when several motifs appear at different network layers in multi-layered network structures.

References

1. Arenas, A., Fernández, A., Fortunato, S., Gómez, S.: Motif-based communities in complex network. J. Phys. A: Math. Theor. **42**(22), 224001 (2008)
2. Benson, A.R., Gleich, D.F., Leskovec, J.: Tensor spectral clustering for partitioning higher-order network structures. In: Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, 30 April–2 May 2015, pp. 118–126 (2015)
3. Benson, A.R., Gleich, D.F., Leskovec, J.: Higher-order organization of complex networks. Science **353**(6295), 163–166 (2016)
4. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefevre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **10**, P10008 (2008)
5. Cover, T.M., Thomas, J.A.: Elements of Inf. Wiley, Theory (1991)
6. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. In: Proceedings of National Academy of Science. USA 1999, pp. 7821–7826 (2002)
7. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)
8. Grünwald, P.D., Myung, I.J., Pitt, M.A.: Advances in Minimum Description Length: Theory and Applications. MIT Press, Cambridge (2005)
9. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
10. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
11. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science **353**(298), 824–827 (2002)
12. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. **E69**, 026113 (2004)
13. Park, Y.J., Song, M.S.: A genetic algorithm for clustering problems. In: Proceedings of 3rd Annual Conference on Genetic Algorithms. Morgan Kaufmann Publishers, pp. 2–9 (1989)
14. Pizzuti, C., Socievole, A.: An evolutionary motifs-based algorithm for community detection. In Proceedings of 8th IEEE International Conference on Information, Intelligences, Systems and Applications (IISA 2017) (2017)
15. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. **105**(4), 1118–1123 (2008)
16. Schaeffer, S.E.: Survey: graph clustering. Comput. Sci. Rev. **1**(1), 27–64 (2007)
17. Serrour, B., Arenas, A., Gómez, S.: Detecting communities of triangles in complex networks using spectral optimization. Comput. Commun. **34**(5), 629–634 (2011)
18. Ulanowicz, R.E., Bondavalli, C., Egnotovich, M.S.: Network analysis of trophic dynamics in South Florida ecosystem, FY 97: the Florida Bay ecosystem. Annual Report to the United States Geological Service Biological Resources Division Ref. No.[UMCES] CBL, pp. 98–123 (1998)