Motif Mining in Weighted Networks

Sarvenaz Choobdar, Pedro Ribeiro, and Fernando Silva
CRACS and INESC-TEC
Faculdade de Ciencias, Universidade do Porto, Portugal
Email:{sarvenaz,pribeiro,fds}@dcc.fc.up.pt

***Abstract*—Unexpectedly frequent subgraphs, known as motifs, can help in characterizing the structure of complex networks. Most of the existing methods for finding motifs are designed for unweighted networks, where only the existence of connection between nodes is considered, and not their strength or capacity. However, in many real world networks, edges contain more information than just simple node connectivity.**

In this paper, we propose a new method to incorporate edge weight information in motif mining. We think of a motif as a subgraph that contains unexpected information, and we define a new significance measurement to assess this subgraph exceptionality. The proposed metric embeds the weight distribution in subgraphs and it is based on weight entropy. We use the g-trie data structure to find instances of k-sized subgraphs and to calculate its significance score. Following a statistical approach, the random entropy of subgraphs is then calculated, avoiding the time consuming step of random network generation.

The discrimination power of the derived motif profile by the proposed method is assessed against the results of the traditional unweighted motifs through a graph classification problem. We use a set of labeled ego networks of co-authorship in the biology and mathematics fields. The new proposed method is shown to be feasible, achieving even slightly better accuracy. Since it does not require the generation of random networks, it is also computationally faster, and because we are able to use the weight information in computing the motif importance, we can avoid converting weighted networks into unweighted ones.

***Index Terms*—Complex Networks, Network Motifs, Weighted networks, Information Theory, Entropy**

I. Introduction

A large body of knowledge has been developed for pattern mining in networks [1], [2], [3], [4], [5], since it has applications in a broad range of fields such as sociology [6], biology [7] or transportation networks [8], where entities are modeled as nodes that are connected if they have interactions or are related. However, most of the developed methods are dedicated to unweighted networks, without taking into account the strength or capacity of the connections.

A pattern in a network is normally defined as a subgraph which is very frequent or infrequent (in case of anomalies). A specific form of patterns are called motifs, which can be thought of as small subgraphs that appear in a network at significantly higher frequencies than what would be expected in similar randomized networks [9]. This type of patterns can really help in characterizing the networks, since they are not frequent only by chance, and therefore really highlight the specific structural properties of the networks. That is why motifs are also known as the building blocks of networks. It has been demonstrated that they can have functional significance in transcriptional regulatory networks [10] or protein-protein interaction networks [11].

We note that network motifs, by their definition, are different from frequent subgraphs [1], [2] or substructures [12]. For an unweighted network with binary connections (where two nodes are either connected or not), motif mining consists essentially in enumerating all subgraphs of specific sizes in a network, and finding those that appear more frequently than expected [9]. This full subgraph enumeration leads to a higher computational complexity for motif mining algorithms, when comparing to frequent subgraph mining algorithms where pruning criteria such as the anti-monotonicity property are used to limit the search space and to improve efficiency. Another restricting issue in motif mining is the calculation of random frequency of subgraphs. From a statistical point of view, the random frequency of a subgraph is reliable only if a reasonable number of random subgraphs are generated for this purpose. These properties imposed by the definition of motifs make it computationally hard to increase the size of subgraphs.

For a better characterization of complex networks, one needs to utilize all available information, including the weights of the edges. This is important in networks such as the traffic flow in a transportation network, strength of social relations, or connectivity strength between every pair of brain regions. To find patterns in weighted networks, the majority of the existing methods need a weight threshold over edges to convert a weighted network to an unweighted one, where nodes are connected if the weight is more than the threshold. A big challenge for this approach is to find an appropriate value for the threshold, and different choices of values lead to very different network topologies. For example, two nodes that are connected in a network for threshold a might be disconnected in a network with threshold b. A limited number of methods were designed to find patterns considering the weight information and to tackle this issue [13], [14]. They propose a weighted support measure for frequent subgraph mining algorithms, based on average weights.

In a weighted network, one requires a measure different from the usual frequency. Saramaki et al. [15] used the average of weights to find motifs in a network. They define two measures, intensity and coherence, based on the average of weights in instances of a particular subgraph type. A subgraph is a motif if these measurements differ from random values.

In this paper we propose a new method to find the subgraphs that are significant in terms of weight distribution. Hence, we need a significance measure that incorporates the weight information. Analogous to motif mining in unweighted networks, a subgraph is relevant if the value of the designed measure is significantly different from its expected random value. We use Shannon's concept of information entropy [16] as the significance measure. Information entropy gives a quantitative measure to assess the amount of latent information in different objects. Entropy measures the uncertainty of a variable; the more randomness it has, the higher the entropy is. Entropy is also used as a measure to differentiate random occurrences or noise in datasets. Given this, it fits well in the problem of motif mining where motifs are the ones which appear in different frequencies than it would be expected in randomized networks. An entropy based approach was also successfully used to discover colored motifs in biological networks [17]. Our approach is however conceptually different, because we incorporate weight information, while this other approach considers unweighted edges and different node classes.

We can say that a subgraph contains almost no information if its weight distribution is completely random. Therefore, a subgraph is relevant, and characteristic of the network, if its weight entropy differs significantly from the weight entropy in random networks. To calculate the random entropy, we exploit an analytical approach. In this way, we greatly decrease needed computation time, avoiding the costly step of having to do an exhaustive random network generation for assessing subgraph significance.

Motif mining methods for unweighted networks mainly fall into two main conceptual approaches. Network-centric methods look for all possible k-sized subgraphs, by enumerating connected sets of k vertices, and in the end they do tests to discover the isomorphic class of each subgraph found. ESU [18] and Kavosh [19] are examples of two state of the art algorithms following this methodology. Subgraph-centric approaches, on the other hand, query individual subgraphs one at the time. Grochow and Kellis [20] developed and efficient algorithm for this.

We have recently developed a new specialized data structure, g-tries[21], that can efficiently represent and query any collection of subgraphs, following an intermediate set-centric approach, in which we really define the custom set of subgraphs we are interested in. G-tries are multiway trees that take advantage of common substructures in the subgraphs to efficiently search at the same time for occurrences of all the subgraphs in the collection. G-tries have been shown to be significantly faster than previous methods when finding motifs [21], [22], and we used them to calculate the more traditional unweighted significance score of subgraphs.

In order to evaluate the feasibility of our new approach, we tackle a graph classification problem, by using a set of subgraphs as a feature vector for classifying networks. We compare the accuracy obtained when using our new weighted motif feature set and the traditional unweighted motif profile. The results show that the proposed measure is feasible and can find a set of subgraphs that help to classify the networks.

The reminder of paper is organized as follows. Section II defines the proposed significance measure. Section III describes our practical implementation of our new metric. Section IV discusses the experimental results and the evaluation of the developed method. Finally, section V concludes and gives possible future directions.

II. Weight entropy of subgraphs

In this section the necessary concepts for formalizing motifs in weighted networks are described and the proposed significance measure, weight entropy, is introduced.

Since edge weights may be continuous values, it is not straightforward to include them in the mining methodology. For a weighted network, we need a measure that not only considers the frequency, but also includes the weight distribution over the edges in a subgraph. In other words, we need a measure that can assess the whole information embedded in a subgraph in order to assign an importance degree for it to be a pattern.

The occurrence of a subgraph is a part of the network characteristics and can describe the functionality or class of the network, if this occurrence does not happen by chance. Hence, if its information content is different from random networks, it is not random. Such a value can be used as a measurement to distinguish the subgraphs in a network as relevant patterns. Information content of an event is commonly used as a discriminating measure [23]. Shannon's theory of information [16] gives us a mathematical tool to quantify the amount of information gained from an event. Information theory assesses how surprising, or unexpected, an observation or event is. If an event always happens, there is no information gain in detecting this event.

Entropy is a function of the probability distribution $P = (p_1, .., p_n)$ where p_i is the probability of occurrence of an event. By defining the occurrence of a subgraph with an edge weight distribution as an event, we can use entropy as a measure to quantify the importance of subgraph for being a pattern. This measure not only considers the weight distribution in the form of probability function, but also assesses the information content of a subgraph.

If X is the random variable describing a particular subgraph g_h^k with k nodes and h edges in a network then it can have different states regarding different edge weight set $\vec{W} = \{w_i \mid i = 1, .., h\}$. The weight entropy of a subgraph is:

$$H_{\vec{W}}(X) = -\int p(X) log(p(X)) \tag{1}$$

where $p(X)$ is the probability of occurrence of weight set $\vec{W}$ in the subgraph g_h^k and is given by:

$$p(X) = P(\vec{W} \leq W) \tag{2}$$

where W is a vector of upper bounds for weights of edges in the subgraph.

For each particular type of subgraph of size k in a network, we assign a weight entropy that reflects the weight distribution in the subgraph and shows if the distribution is random or describes a property in the network.

In this paper, we define a weighted motif as a subgraph whose weight entropy is significantly different from random weight entropy:

$$\left|H_R - H_{\vec{W}}\right| > \delta \tag{3}$$

where H_R is the weight entropy in random networks, called random entropy and δ is a user-defined threshold to find motifs.

An essential step of unweighted motif mining methods is the random simulation for calculating the mean and variance of a subgraph frequency in similar random networks [24], typically conserving the degree sequence of the original networks. This step is computationally very expensive. In this paper, for calculating the random entropy, we do not need this exhaustive generation of random networks, but instead we use analytical formulas to find the probability of occurrence of a subgraph g^k with weight set $\vec{W}$ in an Erdös-Rényi (ER) random graph model. This probability is the main element for calculating the random entropy regarding the equation 1 and is equal to:

$$P_{\vec{W}}^{g_h^k} = p(\vec{W}) * \mu(g^k) \tag{4}$$

where $p(\vec{W})$ is the probability that edges in g_h^k have weight set $\vec{W} = \{w_i \mid i = 1, .., h\}$ and $\mu(g^k)$ is the probability occurrence of a subgraph g_h^k. The first component, denoted by $p(\vec{W})$ follows $f(w)$, the weight distribution in the original network. The joint probability is as follows where the weight of edges in a random network are independent:

$$p(\vec{W}) = P(\{w_i \mid i = 1, .., h\}) = \prod_i^h f(w_i) \tag{5}$$

In a random graph G over a set of nodes V, connectivity between every two nodes i and j is independent and identically distributed in the networks. Edges are described by a set of variables $X = \{X_{i,j}\}$ for all $i, j \in V$ where $X_{i,j}$ is 1 if two nodes are connected, and it is 0 if not. This stationary property of process of random network generation entails that the edge distribution in a network is independent of permutation of nodes, meaning the probability of occurrence of an edge between two nodes i and j does not depend on (i, j) (exchangeable assumption). Picard et al. proposed an analytical method to find the probability of occurrence of a motif in every random network where random variable X is iid [25]. The probability of motif occurrence is independent of the occurrence position. For the ER model, where the exchangeable assumption holds, the probability of occurrence of g^k is as follows:

$$\mu(g^k) = \prod Pr\{(X_{i,j} = 1)\}^{e_{ij}} = \alpha^h \tag{6}$$

where h is the number of edges in g^k and e_{ij} is 1 if nodes j and j are connected and 0 otherwise, for all $i, j \in V(g^k)$. Finally, by substituting the random probability of occurrence subgraph g^k with weight set $\vec{W}$ in formula 1, the random weight entropy is equal to:

$$H^R = -\int_{w_i} \alpha^h * f(w_i)^h log((\alpha * f(w_i))^h) \tag{7}$$

III. Weighted motif mining

We now describe how we implemented our proposed measurement. We are trying to find significant subgraphs, called motifs, having in consideration not only frequency but also weight distribution of edges in the subgraph. Formally, motifs of size k in a weighted network are subgraphs with k nodes where weight distribution has higher entropy than random weighted networks. To implement the weighted method described in section II, we modified the g-tries search algorithm to find such subgraphs and calculate the entropy measure. We divide the implementation into two parts, namely weight entropy of subgraphs in the original network, and random weight entropy for random networks:

A. Entropy of subgraphs in original network

The overall process for finding motifs of size k in a weighted network is that we first need to find all subgraphs of size k (storing the weight set over the edges for each subgraph type i), and secondly we find the probability of occurrence of g_i^k with weight set $\{w_1, w_2, ...w_h\}$. This probability is a multivariate function whose dimension increases as the number of edges in the subgraph increases. To find the weight distribution of a given subgraph, we use the stored weight sets while enumerating the instances of the subgraph in the original network.

We use g-tries [21] for storing and searching for subgraph occurrences. G-tries are multiway trees that are able to store a collection of subgraphs. Their basic principle is to identify common substructure. Subgraphs with the same parent g-trie node share the same topological structure with the exception of a single node and its connections, as is exemplified Figure 1.

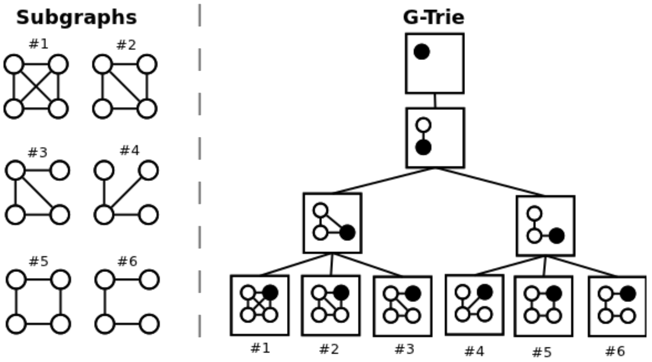

Fig. 1: An example g-trie storing all possible undirected subgraphs of size 6. In each g-trie node, the black vertex is the new one being added, and the white vertices are the ones "inherited" from the parent g-trie nodes.

By using an efficient canonical labeling procedure and symmetry breaking conditions, g-tries allow the search at the same time for an entire set of subgraphs. This avoids the redundancy of searching several times for the same substructure that belongs to different subgraphs, as it would happen if we would search for each subgraph type individually, in a subgraph-centric algorithm such as Grochow and Kellis [20].

At the same time, g-tries also do isomorphism testing as we are traversing the g-trie tree, since when we are at a leaf we can be certain that the subgraph found is of that type. This contrasts with network-centric methods such as ESU [18], which enumerate all connected sets of the desired number of vertices and postpone isomorphism tests to when an entire occurrence is found, not reusing information from previous isomorphisms found.

We modified the original g-tries algorithm so that we are able to store sets of edge weights for each subgraph type, instead of simple integer frequency. After discovering all occurrences of a subgraph g^k in the network, we find its multi-dimensional distribution of weights and calculate H_w of subgraph g^k in the network, regarding equation 1.

B. Random entropy

The second component for finding motifs is the calculation of the random entropy, regarding equation 3. This step is different from conventional motif mining method in the sense that we do it in an analytical way, avoiding the need for generating random networks and computing the frequency of subgraphs in this ensemble of networks. We follow the statistical scenario described in section II.

In calculating the random entropy, the main component is the weight distribution $f(w)$ in the whole network regarding equation 7. An approach to find a distribution is to build the histogram of the data. We use a discretization method to find the histogram, and there are several methods for this purpose. Some of them are supervised methods that need a class label, such as an entropy based method, and others are unsupervised, such as equal width or equal frequency. Here, we use an equal frequency method since we do not have any class label and also because this method finds the intervals that have enough instances for inference, avoiding the generation of sparse intervals in terms of frequency. Equal frequency discretization divides the range of weights for an edge into r intervals where each interval includes n/r values, and n is the number of weight sets. In this way, we have a set of break points $b_1, ..., b_{r-1}$ and a set of frequency counts that define r intervals in the range of each edge weight: $(-\infty, b_1], [b_1, b_2], ..., [b_{r-2}, b_{r-1}], [b_{r-1}, \infty)$. Label b_i is assigned to values belonging to interval i. Finally, the random entropy is calculated using equation 7 and weight distribution $f(w)$.

IV. Evaluation and Results

The main motivation of this paper was to build a method capable of finding relevant subgraphs that characterize the networks. Therefore, we need a way to show that incorporating the weight information in motif mining algorithms can find the right subgraphs as motifs that best represent the functionality or class of the network. One of the main applications of motif discovery using any algorithm is to predict the type of the network. We can build motif profiles that can be used as fingerprints for network classification in different domains such as biological [26] or social [27] networks. Given this, as an evaluation method, we decided to use a classification problem where there is a set of networks with pre-defined labels or classes and the motif profile is used as a feature vector for classification.

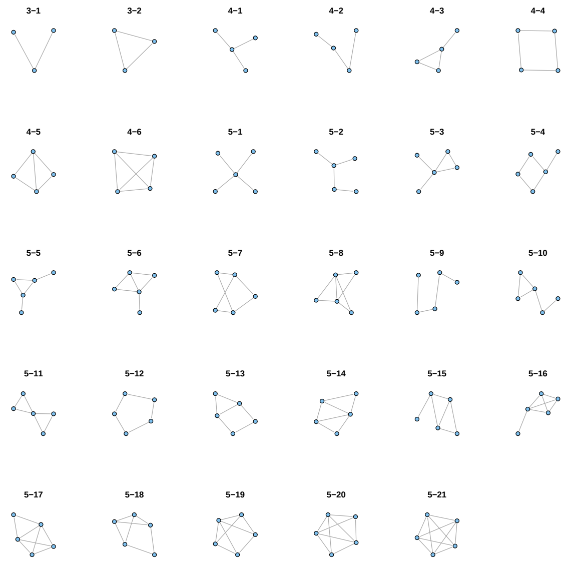

Fig. 2: Set of subgraphs used for motif mining in the ego networks.

For this purpose, we need a dataset of labeled networks. In this paper we use the co-authorship networks of publications authored by researchers from the University of Porto, ranging from 2003 to 2011. These are publications drawn from ISI Thompson Web of Knowledge. We randomly selected 100 authors from two different scientific fields: biology and mathematics. Then, for each author, we built the ego net of authors' collaborations, that is, the network composed solely by the authors that have at least one paper co-authored with him, and their respective interconnections (co-authorship of papers). The label of each ego network is the scientific field that the author belongs to. We selected 30% of authors from mathematics and the others from biology. The weight of the edges is the number of papers that two authors published together.

We apply our proposed method and also the more classical unweighted version of motifs on the dataset to derive the motif profiles which are then used as a feature vector for a standard classifier. Then, the accuracy of classification using both the weighted and unweighted methods are compared to assess the obtained performance in finding the correct motifs in the networks.

We use a variety of classification techniques for the evaluation, including: (i) Decision Trees (C4.5) [28], (ii) Naive Bayesian Classifiers (NB) [29], and (iii) Support Vector Machines (SVM) [30]. The classification results were computed using 10-fold cross validation.

The proposed method and the unweighted one both generate a vector of importance values for subgraphs, respectively called h-score score and z-score. In an unweighted network, the significance of a subgraph is measured in terms of a z-score:

$$z\text{-}score_k = \frac{freq_{original}(G_k) - \overline{freq_{random}(G_k)}}{\sigma(freq_{random}(G_k))}$$

where $\overline{freq_{random}}$ and $\sigma(freq_{random})$ are respectively the average and standard deviation of the frequency in the randomized networks. We derived the motif profile of networks for subgraphs of size 3 to 5 (the usual size in motif mining studies), depicted in Figure 2.

Figures 3 and 4 depict the kernel density estimates of importance scores for the used 100 ego networks in biology and mathematics fields. The plots give the probability that the score of a subgraph fall in an interval. Although both measures give very similar results, h-score values are more concrete and less scattered. As we can see in the figures, h-scores of subgraphs are more centralized around the mean value of importance measure. Hence, if a subgraph is a promising feature in a network, h-score tends to give a stronger value to it. In the figures, the green baselines show the threshold of ±0.6. Regarding the baselines, we can see that if a subgraph is a motif the h-score can detect it with higher probability than z-score. Comparing the histograms across the two research fields, biology and mathematics, we can see clearly that both measures give higher score to different sets of subgraph for each field. For example, subgraphs of size 5 have higher average importance in biology, specially subgraph 5-20 and 5-21 which are more connected. While in mathematics, the average score for smaller and less connected subgraphs, such as 4-1 and 5-1, is higher. The observed pattern for these two fields are in good accordance with results derived in our previous work [27] where co-authorship networks are compared across different scientific fields by their motif profile.

We did the classification with two different scenarios: binary feature vectors and continuous values. In the first scenario, we have a binary vector of size 29 (the number of subgraph types) where we use 1 if the importance value of subgraph is above a defined threshold δ, and we use 0 if it is below the threshold. In the second scenario, we used the original value of motif profiles. For the purpose of comparison we normalize the significance values of both methods into interval of $[-1, 1]$.

The accuracy of built models for the two motif mining methods are compared in table I. The last row of the table shows the results for the case in which we used the continuous values of motif profiles.

From table I, we can see that both methods achieve reasonably good results. Compared to the unweighted version of motifs, the proposed method, not only can characterize the networks, but can also do it with slightly better accuracy. In addition, it has two advantages. First, it takes advantage of weight information in the networks and there is no need for putting a threshold over the weight of edges. Secondly, since this method mainly relies on the distribution of weight in the network, we could use statistical methods to calculate the random value of entropy and we avoid the expensive computational step of motif mining algorithm, which is the random network simulation and correspondent subgraph frequency computation, for measuring the z-score.

TABLE I: The accuracy of the classifiers using the new proposed weighted motif mining method and the classical unweighted method.

threshold δ	weighted motifs			unweighted motifs		
	C4.5	NB	SVM	C4.5	NB	SVM
0.2	80.7	74.1	69.3	71.2	69.4	64.2
0.4	79.9	72.7	72.3	76.5	73.6	67.8
0.6	81.2	75.3	71.3	82.1	75.4	68.1
(continuous)	71.9	68.3	64.3	65.7	67.8	61.7

V. Conclusions

Many real complex networks contain more connectivity information than a simple boolean function that tells us whether a pair of nodes is connected or not. The edges can have weights that greatly improve the expressiveness and information content of the connections. For instance, on co-autorship networks, an unweighted network would not distinguish a connection between two authors that wrote dozens of papers together from a connection between a pair of authors that only were co-authors on a single paper. The same concept can be applied in many other network types, expressing for example the amount of traffic flow in a transportation network, or the connectivity strength between brain regions. In this paper we proposed precisely a novel methodology that is able to find motifs in weighted networks, incorporating the weight information in its calculations.

It is has been shown that subgraph patterns, or motifs, can characterize the functionality of unweighted networks [26]. We defined motifs in weighted networks as the subgraphs that include unexpected information content, that is, that are different from random networks. We proposed a new significance measure based on weight entropy of subgraphs. In our method, we exploit an analytical approach instead of random networks generation for calculating random entropy.

The derived results are compared against unweighted motifs in terms of capability for network characterization. With this purpose in mind, a graph classification problem is used to evaluate the results. The evaluation shows that the proposed method is able to find the set of subgraphs that can differentiate networks at least as well as unweighted motifs, achieving even slightly better accuracy. However, our method is even faster to compute, given that we avoid the random network generation.

In the future we intend to apply our methodology to other networks, exploring its power for general characterization of any type of complex networks. We also intend to use a more broad set of subgraph types (for instance using larger sizes) and to experiment with other random network models, different from the used Erdös-Rényi model.

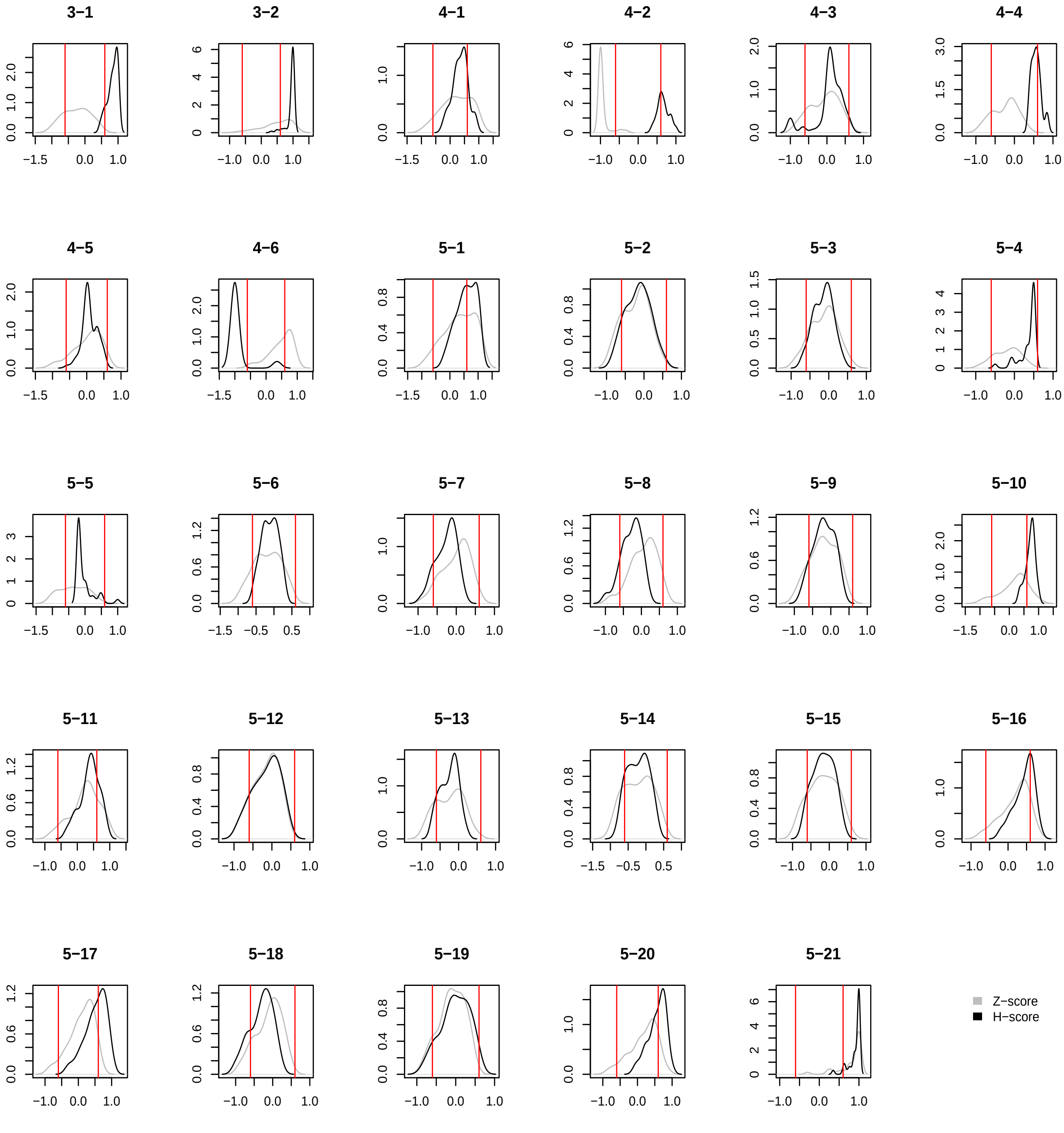

Fig. 3: Kernel density estimate of significance scores, h-score and z-score, for subgraph size 3-5 for biology ego networks. The red vertical base lines depict the threshold of ±0.6 to consider a subgraph as a motif.

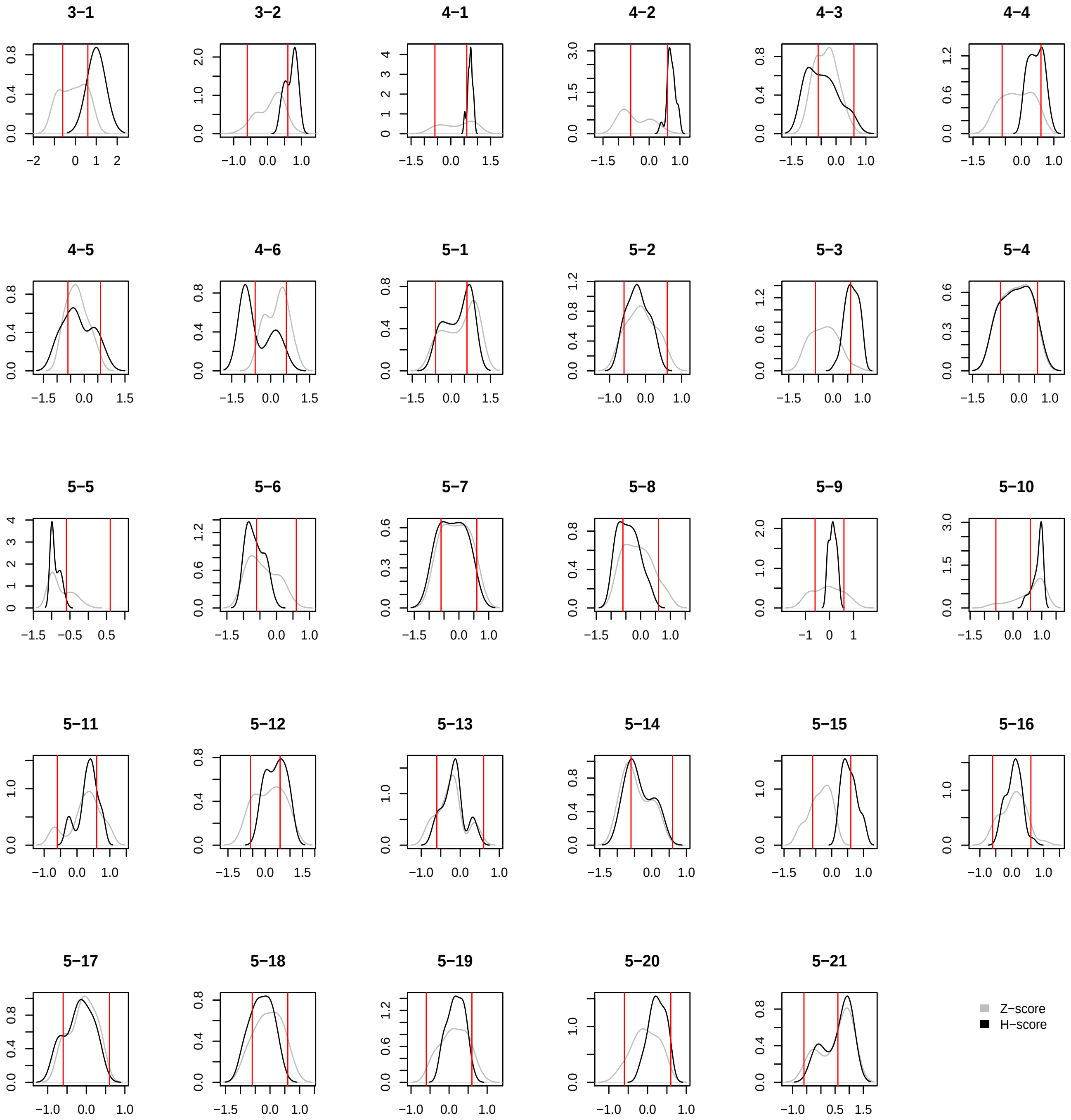

Fig. 4: Kernel density estimate of significance scores, h-score and z-score, for subgraph size 3-5 for mathematics ego networks. The red vertical base lines depict the threshold of ±0.6 to consider a subgraph as a motif.

Acknowledgments

We very thankful to the anonymous referees for their valuable comments. This work is in part funded by the ERDF/COMPETE Programme and by FCT within project FCOMP-01-0124-FEDER-022701. Sarvenaz Choobdar is funded by an FCT Research Grant (SFRH/BD/72697/2010). Pedro Ribeiro is funded by an FCT Research Grant(SFRH/BPD/81695/2011).

References

[1] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *Proceedings of the 2002 IEEE International Conference on Data Mining*, ser. ICDM '02, 2002, pp. 721–.

[2] J. Huan, W. Wang, and J. Prins, "Efficient mining of frequent subgraphs in the presence of isomorphism," in *Proceedings of the Third IEEE International Conference on Data Mining*, ser. ICDM '03, 2003, pp. 549–.

[3] A. Inokuchi, "Mining generalized substructures from a set of labeled graphs," in *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, 2004, pp. 415–418.

[4] 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.

[5] M. Kuramochi and G. Karypis, "Grew-a scalable frequent subgraph discovery algorithm," in *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 2004, pp. 439–442.

[6] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories*. ACM, 2006, pp. 137–143.

[7] M. Koyutürk, A. Grama, and W. Szpankowski, "An efficient algorithm for detecting frequent subgraphs in biological networks," *Bioinformatics*, vol. 20, no. suppl 1, pp. i200–i207, 2004.

[8] W. Jiang, J. Vaidya, Z. Balaporia, C. Clifton, and B. Banich, "Knowledge discovery from transportation network data," in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. IEEE, 2005, pp. 1061–1072.

[9] 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.

[10] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[11] I. Albert and R. Albert, "Conserved network motifs allow protein–protein interaction prediction," *Bioinformatics*, vol. 20, no. 18, pp. 3346–3352, 2004.

[12] C. Helma, S. Kramer, and L. De Raedt, "The molecular feature miner molfea," in *Proceedings of the Beilstein-Institut Workshop*. May, 2002.

[13] C. Jiang, F. Coenen, and M. Zito, "Frequent sub-graph mining on edge weighted graphs," *Data Warehousing and Knowledge Discovery*, pp. 77–88, 2010.

[14] F. Eichinger, K. Böhm, and M. Huber, "Mining edge-weighted call graphs to localise software bugs," *Machine Learning and Knowledge Discovery in Databases*, pp. 333–348, 2008.

[15] J. Saramaki, J.-P. Onnela, J. Kertesz, and K. Kaski, "Characterizing motifs in weighted complex networks," *AIP Conference Proceedings*, vol. 776, no. 1, pp. 108–117, 2005.

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

[17] C. Adami, J. Qian, M. Rupp, and A. Hintze, "Information content of colored motifs in complex networks," *Artificial Life*, vol. 17, no. 4, pp. 375–390, 2011.

[18] S. Wernicke, "Efficient detection of network motifs," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 3, no. 4, pp. 347–359, 2006.

[19] Z. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E. 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.

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

[21] P. Ribeiro and F. Silva, "G-tries: an efficient data structure for discovering network motifs," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, 2010, pp. 1559–1566.

[22] P. Ribeiro and F. Silva, "Querying subgraph sets with g-tries," in *2nd ACM SIGMOD Workshop on Databases and Social Networks (DBSocial)*, 2012, pp. 25–30.

[23] P. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 32–41.

[24] P. Ribeiro, F. Silva, and M. Kaiser, "Strategies for network motifs discovery," in *e-Science, 2009. e-Science'09. Fifth IEEE International Conference on*. IEEE, 2009, pp. 80–87.

[25] F. Picard, 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.

[26] 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.

[27] S. Choobdar, P. Ribeiro, S. Bulga, and F. Silva, "Co-authorship network comparison across research fields using motifs," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[28] J. Quinlan, *C4. 5: programs for machine learning*, 1993.

[29] T. Mitchell, "Machine learning," in *McGraw Hill*, 1996.

[30] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.