Recurrent structural motifs reflect characteristics of distinct networks

*Chen-Hsiang Yeang[1], Liang-Cheng Huang[1,2], Wei-Chung Liu[1]
[1]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
[2]Department of Information Engineering, National Taiwan University, Taipei, Taiwan
*Corresponding author; address: 128 Academia Road, Section, 2, Nankang 115, Taipei, Taiwan
tel: +886-2-27835611; email: chyeang@stat.sinica.edu.tw

***Abstract*—In large-scale networks, certain topological patterns may occur more frequently than expected from a null model that preserves global (such as the density of the graph) and local (such as the connectivity of each node) properties of the graph. These *network motifs* are the building blocks of large-scale networks and may confer functional/mechanistic implications of their underlying processes. Despite active investigations and rich literature in systems biology, network motifs are less explored in social network studies. In this work, we modified and improved the method from Milo et al. 2002 to detect significantly enriched motifs in both directed and undirected networks. We applied this method to identify 3-node and 4-node motifs from the datasets of 18 networks (4 directed and 14 undirected) covering social interactions, co-authorships, web document hyperlinks, neuronal circuitry, protein-protein interactions (PPI), trophic relations in a food web, and others. Presence and absence of enriched motifs provide rich information regarding each type of network relations. In undirected networks, triangles are enriched in almost all datasets, suggesting the prevalence of transitivity in diverse networks. However, 4-node structures lacking transitivity – diamonds and stars – are also enriched in the majority of undirected networks. In directed networks, variations of feed-forward loops are over-represented in the networks of web document and political weblog hyperlinks as well as neuronal connections. In contrast, the food web is enriched with unidirectional motifs with distinct trophic levels. These results reveal the nature of distinct types of networks and invite further explorations on the relations of network structures and types of relations.**

***Keywords*-network motifs; directed graph; undirected graph; permutation tests;**

I. Introduction

Large-scale networks have ubiquitous presence in a wide range of phenomena such as communication connections, World Wide Web document hyperlinks, social interactions, gene regulatory circuitry, and ecological webs. The sheer size of the networks, inter-dependencies between the constituting members, and diverse types of possibly encoded relations make the analysis of network data challenging. Investigations on large-scale networks are roughly categorized into two classes. Intrinsic studies probe the characteristics of the networks and attempt to find general properties beyond individual application domains or the properties pertaining to specific phenomena. Instances include discoveries of the "scale-free" and "small-world" properties of large-scale networks [1], [2], [3], detection of communities in social networks [4], [5], [6], propagation of culture/opinions in social networks [7], and relations between complexity and stability of ecological networks [8]. Extrinsic studies exploit external information in addition to network structures and attempt to employ the information of network structures to understand/predict external features or vice versa. Instances include designing network structures to facilitate collaboration on certain tasks [9], [10], applying social/cultural/genetic features to explain formation of friendships [11], and predicting consumer preferences based on their social networks [12].

One of the most important topics in intrinsic studies is network motif detection. A motif is a recurrent pattern that appears in multiple instantiations of an application domain, such as biology (protein structural motifs and DNA sequence motifs), music (music ideas of a symphony), literature (recurring narrative elements in a story), and visual arts (a graphical pattern that populates in an Islamic textile artwork). These recurrent patterns are considered as building blocks of entities (a protein, the promoter of a gene, a symphony, a tapestry), and different combinations of a small number of motifs can generate enormously diverse forms. Similarly, a network motif is a graph structure (topology) that recurs in multiple parts of a large-scale network. The concept of network motifs was first coined by Alon in systems biology [20]. His team devised statistical methods to detect network motifs with significant enrichment [13]. Moreover, they provided functional arguments to support enrichment of certain network motifs in gene regulatory networks [14], and experimentally validated these arguments in selected cases [15]. Since then network motif detection has become a sub-discipline in systems biology.

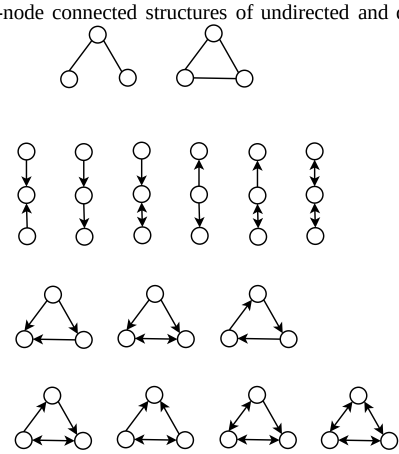
In the arena of social networks, sociologists have intensively studied small network structures – particularly diads (two-person interactions) and triads (three-person interactions) – for a long time (e.g., [16], [17]). Faust counted the occurrences of all possible triads in a variety of social networks and used the occurrence count vectors to categorize the interactions encoded by those social networks [18]. However, to our knowledge systematic approaches to identify statistically enriched motifs with more than 3 nodes have not be pursued in social network analysis.

In this work, we modify and improve the method from [13] to detect significantly enriched motifs in both directed and undirected networks. We apply our method to detect 3-node and 4-node motifs from the datasets of 18 networks (4 directed and 14 undirected) covering social interactions, co-authorships, web document hyperlinks, neuronal circuitry, protein-protein interactions, trophic interactions in a food web, and others. We first describe the concept of network motifs and the algorithm to exactly count network motifs in a graph. We then introduce a null model of generating randomized networks that retain key characteristics – node connectivity and the counts of lower-order motifs – of the empirical network. Motifs significantly over-represented in the empirical network according to the null model are reported. Finally, we discuss the functional/mechanistic implications of the presence/absence of enriched motifs.

Table I
NUMBERS OF CONNECTED NETWORK MOTIFS WITH 3-5 NODES

type	3 nodes	4 nodes	5 nodes
undirected	2	6	21
directed	13	199	9364

Figure 1. 3-node connected structures of undirected and directed graphs



II. METHODS

A. Network motifs

Qualitatively, a network motif is a subgraph structure embodying multiple instantiations in a network. In graph theory, an *isomorphism* of graphs G and H is a bijection (one-to-one and onto mapping) between their vertex sets.

$$f : V(G) \to V(H).$$

such that any node pair $u, v \in V(G)$ are adjacent if and only if $f(u), f(v) \in V(H)$ are also adjacent. In other words, two graphs G and H are isomorphic if there exists a bijection mapping f such that the adjacency matrices of $V(G)$ and $f(V(G))$ are identical. Isomorphism applies to both directed and undirected graphs.

All possible graphs with a fixed number of nodes are partitioned into equivalence classes according to isomorphisms. Graphs in the same equivalence class are isomorphic to each other, thus are considered as belonging to the same graph structure. There are far fewer graph structures (equivalence classes) than the possible graphs for a fixed graph size. Table I lists the numbers of connected graph structures with node sizes ranging from 3 to 5. Figure 1 displays all the 3-node, connected structures of directed and undirected graphs. We define a network motif as an equivalence class according to graph isomorphisms.

B. Exact counting of network motifs

As the name suggests, a network motif is a small graph structure (e.g., a triangle) that possesses multiple instantiations in a much larger base network (e.g., the social network of all students in a school or the interaction network of all human proteins). To detect these recurrent motifs it is necessary to count the motif occurrences in a base network. The upper bound of motif occurrences is $O(n^m)$, where n is the number of nodes in the base network and m is the number of nodes in the motif. For relatively small base networks ($n \leq 1000$) and motifs ($m \leq 4$), we propose the decision-tree algorithm in Figure 2 to efficiently count the exact number of motif occurrences. Briefly, we consider all the m-node connected subgraphs of the base network and count the total number of those subgraphs isomorphic to each motif. The root of the decision tree is an empty node, and its children include all nodes in the base network. The children of each non-empty node are its neighbors in the base network, excluding its ancestors in the decision tree. The depth of the decision tree is the motif size (m). A path from the root to a leaf corresponds to an m-node connected subgraph. To avoid double-counting, we only consider one traversing order among all the paths sharing the same nodes.

Figure 2. An exact motif counting algorithm

Inputs: A base network G with n nodes; a complete collection of network motifs M with m nodes.
Outputs: $C(G, \mu)$, the total number of instances in G of each motif $\mu \in M$.
Procedures:

1) Set $C(G, \mu) = 0$ for each $\mu \in M$.
2) For each $\mu \in M$ with each permutation order of node indices, select an order of traversing the decision tree closest to the monotonically increasing order of node indices. For instance, the selected traversing order of the chain $1-2-3$ is 123, while the selected traversing order of $2-3-1$ is 132.
3) Construct a decision tree T with the following procedures.
   a) Create a root node r with an empty label $L(r) = \phi$.
   b) Create n children of the root with labels ranging from 1 to n.
   c) For each internal node ν in T, find its corresponding node (label) $L(\nu)$ in G. Find the neighbors $\mathcal{N}(L(\nu))$ of $L(\nu)$ in G, and exclude the labels of the ancestors of ν in T from $\mathcal{N}(L(\nu))$. The remaining labels constitute the children of ν in T.
   d) If the path from r to ν in T contains $m+1$ nodes or ν has no more valid neighbors, then stop generating children and treat ν as a leaf. Otherwise descend to each child and continue the iteration.
4) Iteratively traverse all paths from r to leaves in T.
   a) If the depth of a leaf node ν is $m+1$ and the traversing order of the path from r to ν is compatible with the selected traversing order for the corresponding network motif and permutation order, then it is a valid path.
   b) For each valid path π in T, obtain the subgraph G_π in G spanned by nodes in π. Find the motif $\mu \in M$ isomorphic to G_π. Increment $C(G, \mu)$ by 1.

Proposition: $C(G, \mu)$ obtained from Figure 2 is the exact count of motif instantiations.

Proof: It suffices to show that there is a bijection between all valid paths in T and all connected m-node subgraphs in G.

The unique mapping from each valid path in T to a connected m-node subgraph in G is obvious. Each valid path constitutes m nodes, and each node along a path is a neighbor of its precedent nodes by construction. Thus nodes on each path are connected.

To demonstrate the unique mapping from each connected m-node subgraph in G to a valid path in T, we need to show that

Table II
RUNNING TIME COMPARISON BETWEEN BRUTE-FORCE AND DECISION-TREE ALGORITHMS OF MOTIF COUNTING. t_1: 3-NODE MOTIFS, DECISION-TREE. t_2: 3-NODE MOTIFS, BRUTE-FORCE. t_3: 4-NODE MOTIFS, DECISION-TREE. t_4: 4-NODE MOTIFS, BRUTE-FORCE.

dataset	# nodes	t_1	t_2	t_3	t_4
word adjacency	112	< 1	1	5	8
karate club	35	< 1	< 1	1	< 1
les miserables	77	< 1	< 1	1	2
dolphin network	62	< 1	< 1	< 1	1
football tournaments	115	< 1	< 1	3	10
political books	105	< 1	< 1	3	7
C. elegans brain	297	< 1	3	273	576
food web	221	1	< 1	33	111

(1)each connected m-node subgraph in G corresponds to at least one valid path in T, (2)valid paths in T are unique. Since the decision tree T covers all possible orders of traversing m connected nodes, each connected m-node subgraph in G corresponds to at least one path in T. Among all the paths traversing the same m-node subgraph, only the one compatible with the selected traversing order is labeled as valid. Hence valid paths in T are unique and cover all m-node subgraphs in G. Q.E.D.

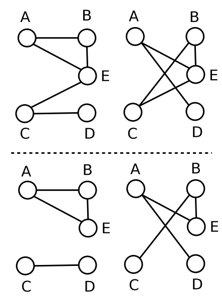
The time complexity of the exact counting algorithm is $O(nd^{m-1})$, where d is the maximum connectivity of nodes in G. It is much smaller than that of the brute-force approach $O(n^m)$ on sparse graphs where $d \ll n$. To demonstrate the utility of the exact counting algorithm, we extracted 7 networks from empirical datasets and applied the brute-force and decision-tree algorithms to count all 3 and 4 node motifs. In all the cases, the two algorithms gave rise to the same counts. Table II shows the running times of the two algorithms. The decision-tree algorithm outperforms the brute-force algorithm in all but one case (karate club, 4-node motifs). The small running time difference in this case is attributed to the small network size: the overhead of enumerating the traversing orders for all motifs and node index orders exceeds the time of exhausting all subgraphs in this small network. The merit of the decision-tree algorithm is more pronounced as the network size increases. For instance, on the food web network with 221 nodes, brute-force counting of 4-node motifs takes 111 seconds, whereas the decision-tree algorithm takes only 33 seconds.

C. Evaluating statistical significance of network motif frequencies

A fundamental hypothesis for the existence of network motifs is that certain graph structures (topologies) confer functional/mechanistic implications thus are over-represented in the base networks. A classical example is the enrichment of feed-forward loops in the gene regulatory networks of *Escherichia coli* [14]. Mangan and Alon argued that feed-forward loops possessed the advantages of stability and quick response times compared to simple linear circuits, thus were evolutionary favorable in the gene regulatory networks.

To detect such functionally/mechanistically favorable network motifs, it is necessary to show that certain network structures are over-represented in the base networks, conditioned on a proper null hypothesis regarding the motif distributions in the base networks. The enrichment results depend on the choice of the null hypothesis. An ideal null hypothesis should capture the characteristic of the base network and exclude the information of the target motif

Figure 3. Top: a subgraph where an edge swap $\overline{AB}, \overline{CD} \to \overline{AD}, \overline{BC}$ preserves the counts of 3-node motifs. Bottom: a subgraph where an edge swap $\overline{AB}, \overline{CD} \to \overline{AD}, \overline{BC}$ annihilates a triangle and creates two chains.



frequencies. Two common null models were previously employed in a wide range of publications. The first null model controls the size and average density ($\frac{\text{\# edges}}{\text{\# node pairs}}$) of the network and generates random Erdös-Rényi graphs accordingly (e.g., [20]). This null model retains the global characteristic of the base network (size and density) but distorts the properties of individual nodes (e.g., connectivity of each node). The second null model randomly swaps edge pairs from the base network (e.g., [19]). A swap on an edge pair with distinct nodes (e.g., converting $\overline{AB}$ and $\overline{CD}$ into $\overline{AD}$ and $\overline{CB}$) does not alter the connectivity of any node. Hence the perturbed networks retain the global (network size, average density, degree distribution) as well as local (connectivity of individual nodes) properties of the base network.

Despite the advantages of the second null model, edge swaps alone may not preserve the statistics of lower-order network motifs from the base network. To claim over-representation of an m-node network motif, we have to exclude the contributions from enrichment of network motifs with less than m nodes. For instance, in networks where triangles are over-represented, the probability of observing any 4-node motif containing a triangle is also elevated compared to random Erdös-Rényi networks or edge-swapped networks. Yet this enrichment is due to over-representation of triangles instead of the true effects of the 4-node motif.

To incorporate the effects of lower-order motif statistics, we consider only the edge swaps that preserve the numbers of lower-order motifs. Figure 3 gives examples of one valid edge swap and one invalid edge swap. In the top diagram, the subgraph contains 1 triangle and 2 chains before and after the edge swap. In the bottom diagram, however, the subgraph contains 1 triangle and no chain before the edge swap and no triangle and 2 chains afterwards. In Figure 4, we describe a method to test whether an edge swap preserves occurrences of lower-order motifs.

Notice when $m = 3$, edge swaps alone preserve the counts of 2-node connected motifs (edges). Hence the lower motif preservation criteria in Figure 4 are not needed. Moreover, the edge swap and validity test apply to both directed and undirected graphs.

Proposition: The edge swaps passing the test in Figure 4 preserve the counts of $(m-1)$-node motifs.

Proof: It suffices to show that only subgraphs obtained from the first step of the test can possibly alter the counts of $(m-1)$-node motifs. Since the only changes are the swapped edges $((v_1, v_2), (v_3, v_4)$ before the edge swap and $(v_1, v_4), (v_3, v_2)$ after

Figure 4. A test verifying whether an edge swap operation preserves occurrences of $(m-1)$-node motifs

Input: A graph G, network motif size $m > 3$, edge pair $(v_1, v_2), (v_3, v_4)$.

Output: Whether edge swap $(v_1, v_2), (v_3, v_4) \rightarrow (v_1, v_4), (v_3, v_2)$ preserves the numbers of $(m-1)$-node motifs in G.

Procedures:

1) Exhaust all subgraphs containing v_1, v_2, v_3, v_4 and k additional nodes, where $k \leq m-3$. Moreover, the k additional nodes are connected to at least one of v_1, v_2, v_3, v_4 with distance $\leq m-3$.
2) For each of those subgraphs, count the number of each $(m-1)$-node motif before and after the edge swap.
3) If there exists at least one subgraph and $(m-1)$-node motif such that the motif counts before and after the edge swap on the subgraph are different, then the edge swap does not preserve the counts of $(m-1)$-node motifs. Otherwise the edge swap preserves the counts of $(m-1)$-node motifs.

the edge swap), only subgraphs containing at least one swapped edge can possibly alter the counts of $(m-1)$-node motifs. Each such subgraph contains $(m-1)$ nodes, and at least two nodes belong to $\{v_1, v_2, v_3, v_4\}$. Hence there are at most $m-3$ additional nodes. Furthermore, only the nodes within distance $m-3$ from $\{v_1, v_2, v_3, v_4\}$ can be possibly contained in the connected $(m-1)$-node motifs. Q.E.D.

To evaluate the statistical significance of an m-node network motif, we perturb the base network by performing valid edge swaps until all the nodes containing swappable edges have been swapped at once. 100 perturbed networks are generated, and the p-value is the fraction of perturbed networks whose motif counts exceed those of the empirical values.

In large and dense networks containing more than 1000 nodes and/or high connected nodes, exact counting of the entire networks is either intractable or time-consuming. For each perturbed network, we identify the nodes undergoing swap operations and extract the subgraphs spanned by the swapped nodes in the empirical and perturbed networks. We then count network motifs in the subgraphs of the empirical and perturbed networks and evaluate the p-value accordingly. This counting is approximate as the non-swapped nodes may also contribute to the changes of motif count differences. Yet the motif counts in the selected subgraphs can faithfully reflect the order (not magnitude) of motif abundance between the empirical and perturbed networks. Thus the approximate counting would not drastically distort the true p-values.

Notice that our method shares the same goal as the one proposed by Milo et al.: to preserve the counts of lower-order motifs in the randomized graphs [13]. Milo et al. defined an energy function as the normalized difference between the counts of lower-order motifs in the empirical and randomized networks. They applied Metropolis Monte-Carlo sampling to generate randomized networks that minimized the energy function. Despite the relative simplicity of implementing the Metropolis Monte-Carlo sampling, the burn-in period for achieving zero energy and the nodes undergoing edge swaps are not explicitly controlled. In these regards our method of explicitly generating valid edge swaps is superior to the Metropolis sampling.

Table III
NETWORK DATASETS ANALYZED IN THIS STUDY. * INDICATES DIRECTED NETWORKS.

abbreviation	description	# nodes/# edges
web-NotreDame*	web document hyperlinks	325729/1497134
celegansneural*	*C. elegans* neuronal network	297/2359
polblogs*	political weblog hyperlinks	1493/19091
foodweb*	Bridge Brook food web	221/553
CA-HepTh	co-authorships on high energy physics	9877/51971
CA-GrQc	co-authorships on general relativity	5242/28980
CA-CondMat	co-authorships on condensed matter physics	23133/186936
netscience	co-authorships on network science	1589/2742
adjnoun	word adjacency in *David Copperfield*	112/425
karate	karate club social network	34/78
lesmis	character co-appearance	77/254
dolphin	dolphin social network	62/159
football	football tournaments	115/615
polbooks	co-purchasing relations of US political books	105/441
Hsapi20101010CR	human protein-protein interactions, core set	1715/1875
Hsapi20101010	human protein-protein interactions, full set	7389/134842
Scere20101010	yeast protein-protein interactions	5212/281810
PTT	PTT user interactions	39752/77970

III. RESULTS

A. Data sources

We extracted 18 network datasets compiled from the webpage http://www-personal.umich.edu/∼mejn/netdata/ and additional sources. Table III summarizes the general information of these networks. They include directed networks of hyperlinks between World-Wide-Web documents (web-NotreDame, [21]), neuronal connections of the worm *Caenorhabditis elegans* (celegansneural, [3]), hyperlinks between weblogs on US politics (polblogs, [22]), and a food web from the Bridge Brook Lake (foodweb, [23]); undirected networks of co-authorships in high energy physics (CA-HepTh, [24]), general relativity (CA-GrQc, [24]), condensed matter physics (CA-CondMat, [24]), and network science (netscience, [25]), undirected networks of word adjacency of common adjectives and nouns in the novel *David Copperfield* (adjnoun, [25]) and co-appearances of characters in the novel *Les Miserables* (lesmis, [27]), undirected social networks between 34 members of a karate club (karate, [26]) and 62 dolphins (dolphins, [28]), undirected networks of American football games between Division IA colleges (football, [4]), co-purchasing relations of books about US politics from Amazon.com (polbooks, [29]), undirected networks of protein-protein interactions in human (core set Hsapi20101010CR, full set Hsapi20101010, [30]) as well as the budding yeast (Scere20101010, [30]), and an undirected network of user interactions in an electronic bulletin board in the PTT electronic board in Taiwan [31]. In PTT users post articles on specified categories (boards), and other users respond to the posted articles. We created an undirected network of PTT users from the records in May 2010. An edge denotes that two users have responded to each other's posted articles.

B. Enriched motifs of undirected networks

Figure 5 displays the enriched 3-node and 4-node motifs in the 14 undirected networks, and Table IV shows their summary statistics and p-values. There are only two 3-node, undirected and connected network motifs: a 3-node chain (motif 0) and a triangle (motif 1). Despite the overwhelming presence of chains, they are not significantly enriched in any the 14 undirected networks. In contrast, triangles are significantly enriched in almost all of the 14 undirected networks. Only two networks – word adjacency and karate networks – have insignificant p-values on triangles (0.23 and 0.1 respectively), while the p-values of all other networks < 0.01 (the occurrences in none of the 100 perturbed networks exceed those in the empirical networks). In addition to significant p-values, the gaps of mean triangle occurrences between empirical and perturbed networks are large in most datasets. In 9 datasets, the mean triangle occurrences in the empirical networks are more than three folds as those in the perturbed networks. An extreme example is the network of co-authorships in network science. The mean triangle occurrence in the empirical network (3764) is 125 times as that in the perturbed networks (30).

Among the six 4-node, undirected and connected motifs, three of them are significantly enriched in at least one dataset. A 3-star (motif 0) has significant p-values ($p \leq 0.05$) in 6 datasets: the co-authorship network in general relativity ($p = 0.03$), the word adjacency network ($p = 0.01$), the character co-appearance network in *Les Miserables* ($p < 0.01$), the football network ($p < 0.01$), the co-purchasing network of political books ($p < 0.01$), and the core network of human protein-protein interactions ($p = 0.02$). In contrast to triangles, the gaps of 3-star mean occurrences between empirical and perturbed networks are not extraordinarily large. Among all the datasets with significant p-values, the mean occurrences in empirical networks are less than two folds of those in perturbed networks.

Diamonds (motif 3) are also highly significantly enriched ($p < 0.01$) in 6 datasets: PTT user interactions, the karate social network, the football network, the co-purchasing network of political books, and both full and core networks of human protein-protein interactions. In all networks, the occurrence frequencies of diamonds are much smaller than those of 3-stars. However, low occurrence frequencies may still confer over-representation if the occurrence frequencies of the perturbed networks are even smaller. For instance, in the PTT dataset the mean occurrences of diamonds are 38 in empirical networks and 17 in perturbed networks. Hence the small number of diamonds are over-represented in the PTT network. In contrast, the mean occurrences of 3-stars are 1586 and 1592 respectively in empirical and perturbed networks. Thus 3-stars are not over-represented in the PTT network despite the higher occurrence frequencies.

The remaining 4-node motifs are either enriched in only one dataset or not enriched at all. A triangle with a dangling edge (motif 2) is enriched only in the football network ($p = 0.02$). A cascade of two triangles (motif 4) has a marginally significant p-value in the full dataset of human protein-protein interactions ($p = 0.06$). A 4-node chain (motif 1) has a marginally significant p-value in the character co-appearance network in *Les Miserables* ($p = 0.07$). Intriguingly, 4-node cliques (motif 5) are not significantly enriched in any undirected network.

C. Enriched motifs of directed networks

Figures 6 and 7 display the enriched 3-node and 4-node motifs in the 4 directed networks, and Tables V and VI show their summary statistics and p-values. Six 3-node motifs containing triangles are significantly enriched ($p \leq 0.05$) in the networks of webpage hyperlinks (web-Notredame), neuronal connections of *C. elegans* (celegansneural), and hyperlinks between weblogs on US politics (polblogs): motifs 7, 6, 9, 10, 11 and 12. Five of these six enriched motifs are compatible with feed-forward loops. Motif 6 is a feed-forward loop with unidirectional links. Motifs 7, 10 and 11 are feed-forward loops with both unidirectional and bidirectional links. Motif 12 is a triangle with bidirectional links only. In contrast, motif 9 contains both unidirectional and bidirectional links and is compatible with a 3-node cycle. The 3-node cycle motif containing only unidirectional links – motif 8 – is enriched only in the network of webpage hyperlinks (web-Notredame, $p = 0.04$). Furthermore, motif 8 has much fewer occurrences in all the 4 directed networks than other trianglular motifs compatible with feed-forward loops. For instance, in the dataset of webpage hyperlinks, the mean occurrences of motif 8 in empirical and perturbed networks are 12 and 3 respectively. In contrast, the mean occurrences of motif 6 in empirical and perturbed networks are 2107 and 1760 respectively.

Three 3-node motifs compatible with chains are enriched only in

Figure 5. Undirected motifs enriched in the datasets

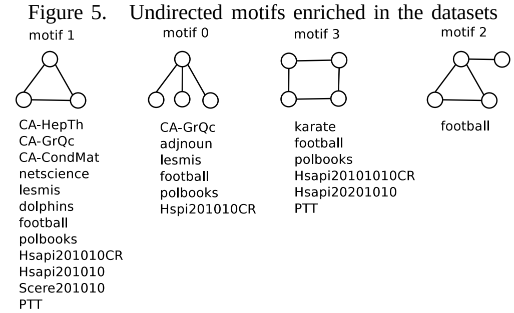


Table IV
SUMMARY INFORMATION OF UNDIRECTED MOTIFS ENRICHED IN THE DATASETS. THE STANDARD DEVIATIONS OF MOTIF COUNTS ARE NOT APPLICABLE WHEN THE MOTIF COUNTS OF THE ENTIRE EMPIRICAL NETWORKS ARE CALCULATED.

# nodes	index	dataset	p-value	empirical mean/std	perturbed mean/std
3	1	CA-HelTh	< 0.01	4341/599	336/643
3	1	CA-GrQc	< 0.01	32016/1756	1616/3820
3	1	CA-CondMat	< 0.01	4156/422	1083/769
3	1	netscience	< 0.01	3764/-	30/6
3	1	adjnoun	0.23	284/-	260/29
3	1	karate	0.1	45/-	29/10
3	1	lesmis	< 0.01	467/-	166/18
3	1	dolphins	< 0.01	95/-	30/6
3	1	football	< 0.01	810/-	149/12
3	1	polbooks	< 0.01	560/-	176/15
3	1	Hsapi201010CR	< 0.01	164/6	4/4
3	1	Hsapi201010	< 0.01	913631/55434	491290/199339
3	1	Scere20101010	< 0.01	1795701/50650	1371269/173110
3	1	PTT	< 0.01	9899/1181	1225/1904
4	0	CA-GrQc	0.03	3108/938	2543/970
4	0	adjnoun	0.01	24781/-	24287/251
4	0	lesmis	< 0.01	6327/-	6245/44
4	0	football	< 0.01	3422/-	3273/39
4	0	polbooks	< 0.01	8128/-	7946/55
4	0	Hsapi201010CR	0.02	4184/499	2956/433
4	3	karate	< 0.01	36/-	23/5
4	3	football	< 0.01	564/-	476/23
4	3	polbooks	< 0.01	557/-	499/21
4	3	Hsapi201010CR	< 0.01	276/28	20/8
4	3	Hsapi201010	< 0.01	16351/3122	2712/1174
4	3	PTT	< 0.01	38/25	17/14
4	2	football	0.02	7536/-	7414/57

the foodweb network: motif 0 (two convergent unidirectional links, $p = 0.01$), motif 1 (a unidirectional 3-node chain, $p = 0.01$), motif 3 (two divergent unidirectional links, $p = 0.01$).

15 4-node motifs are enriched in at least one directed network. Eight motifs are compatible with diamonds: motifs 81, 70, 77, 88, 94, 84, 118 and 131. Motif 81 contains a source, a sink, and two intermediate nodes. The 4 unidirectional links are from the source to intermediate nodes or from intermediate nodes to the sink. It is enriched in all the 4 directed networks. Motif 70 contains unidirectional links connecting 2 sources to 2 sinks. It is enriched in the networks of neuronal connections of *C. elegans*, hyperlinks between weblogs on US politics, and the food web. Motif 88 contains two unidirectional links emitting from a source and two bidirectional links converging to a sink. It is enriched in the networks of neuronal connections of *C. elegans* and hyperlinks between weblogs on US politics. Motif 118 contains two unidirectional links converging to a sink and two bidirectional links emitting from a source. It is enriched in the hyperlinks between weblogs on US politics. Motif 131 differs from motif 81 by replacing one source-intermediate node link and one intermediate node-sink link with bidirectional links. It is enriched in the network of neuronal connections of *C. elegans*. Motif 77 differs from motif 70 by replacing one unidirectional link with a bidirectional link. It is enriched in the networks of neuronal connections of *C. elegans* and hyperlinks between weblogs on US politics.

Three motifs are compatible with diamonds with an additional link: motifs 71, 139 and 82. Motif 71 differs from motif 70 by an additional unidirectional link connecting two sink nodes. It is enriched in the networks of neuronal connections of *C. elegans*, hyperlinks between weblogs on US politics, and the food web. Motif 139 differs from motif 70 by an additional unidirectional link connecting two source nodes. It is enriched in the networks of neuronal connections of *C. elegans* and the food web. Motif 82 differs from motif 81 by an additional unidirectional link between two intermediate nodes. It is enriched in the food web.

Two feed-forward motifs – motifs 94 and 84 – are enriched in the networks of neuronal connections of *C. elegans* (motifs 94 and 84) and hyperlinks between weblogs on US politics (motif 94). The 4-node chain (motif 12) is enriched only in the food web. The motif containing three unidirectional links converging to the common sink (motif 0) is enriched in the hyperlinks between weblogs on US politics. Finally, the unidirectional motif containing two sources, one sink and one intermediate node (motif 1) is enriched in the networks of neuronal connections of *C. elegans* and hyperlinks between weblogs on US politics.

IV. Discussions

The most important conclusion from our analysis is the over-representation of selected local structures in networks. Most subgraph topologies are not enriched in any of the 18 datasets. Existence of network motifs implies that certain structures reflect the intrinsic properties of relations or confer functional importance of the underlying systems, thus are over-represented in the networks. We categorize the 18 networks in this study into several classes according to their enriched network motifs.

For undirected graphs, the majority of co-authorship networks (CA-HepTh, CA-CondMat, netscience) and the dolphin social network (dolphins) are enriched with triangles (3 nodes, motif 1)

Figure 6. 3-node directed motifs enriched in the datasets

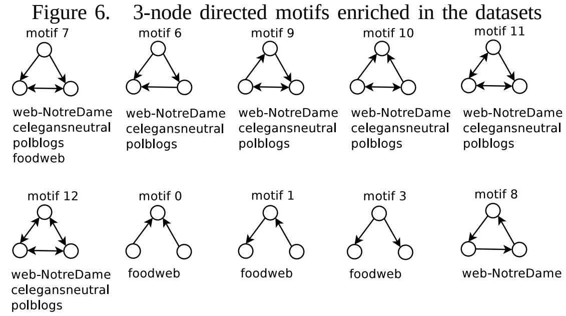


Table V
SUMMARY INFORMATION OF 3-NODE DIRECTED MOTIFS ENRICHED IN THE DATASETS. THE STANDARD DEVIATIONS OF MOTIF COUNTS ARE NOT APPLICABLE WHEN THE MOTIF COUNTS OF THE ENTIRE EMPIRICAL NETWORKS ARE CALCULATED.

# nodes	index	dataset	p-value	empirical mean/std	perturbed mean/std
3	7	web-NotreDame	< 0.01	1092/209	874/175
3	7	celegansneural	< 0.01	312/-	80/9
3	7	polblogs	< 0.01	15436/240	6492/1355
3	7	foodweb	0.05	111/-	77/21
3	6	web-NotreDame	< 0.01	2107/217	1760/196
3	6	celegansneural	< 0.01	1972/-	1427/37
3	6	polblogs	< 0.01	43530/700	27039/1942
3	9	web-NotreDame	< 0.01	43/14	22/9
3	9	celegansneural	< 0.01	179/-	140/11
3	9	polblogs	< 0.01	3827/58	3287/153
3	10	web-NotreDame	< 0.01	7883/1510	6331/1333
3	10	celegansneural	< 0.01	542/-	205/13
3	10	polblogs	< 0.01	14854/226	8774/689
3	11	web-NotreDame	< 0.01	104/31	64/25
3	11	celegansneural	< 0.01	148/-	46/7
3	11	polblogs	< 0.01	10033/141	6032/450
3	12	web-NotreDame	< 0.01	12482/3859	9343/3038
3	12	celegansneural	< 0.01	16/-	3/1
3	12	polblogs	< 0.01	2874/51	1220/218
3	0	foodweb	0.01	6338/-	6017/153
3	1	foodweb	0.01	1931/-	1584/175
3	3	foodweb	0.01	1539/-	1244/139
3	8	web-NotreDame	0.04	12/7	3/3

Figure 7. 4-node directed motifs enriched in the datasets

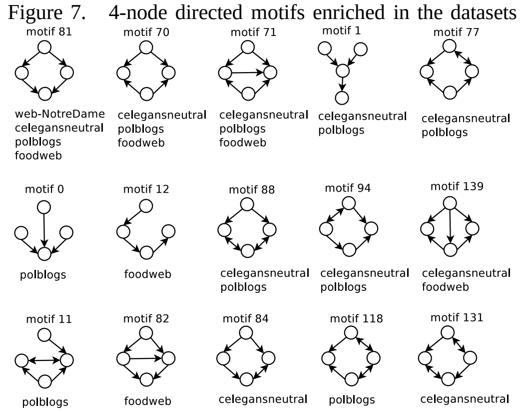


Table VI

SUMMARY INFORMATION OF 4-NODE DIRECTED MOTIFS ENRICHED IN THE DATASETS. THE STANDARD DEVIATIONS OF MOTIF COUNTS ARE NOT APPLICABLE WHEN THE MOTIF COUNTS OF THE ENTIRE EMPIRICAL NETWORKS ARE CALCULATED.

# nodes	index	dataset	p-value	emp mean/std	pert mean/std
4	81	web-NotreDame	0.02	48/28	11/13
4	81	celegansneural	< 0.01	5110/-	4829/50
4	81	polblogs	0.04	1618/392	1546/389
4	81	foodweb	< 0.01	8263/-	7956/110
4	70	celegansneural	< 0.01	3184/-	2820/39
4	70	polblogs	< 0.01	13755/2528	12661/2433
4	70	foodweb	< 0.01	12804/-	12308/118
4	71	celegansneural	< 0.01	2436/-	2354/21
4	71	polblogs	0.03	6895/1950	6603/1880
4	71	foodweb	< 0.01	5073/-	5009/55
4	1	celegansneural	< 0.01	93644/-	93412/67
4	1	polblogs	0.02	80063/16264	73965/15583
4	77	celegansneural	< 0.01	1064/-	998/12
4	77	polblogs	0.01	3025/688	2851/654
4	0	polblogs	0.01	416894/77470	368104/71678
4	12	foodweb	< 0.01	1226/-	1173/48
4	88	celegansneural	0.03	217/-	208/5
4	88	polblogs	0.03	270/79	243/72
4	94	celegansneural	< 0.01	1200/-	1136/19
4	94	polblogs	0.06	1438/372	1383/361
4	139	celegansneural	0.03	1001/-	978/12
4	139	foodweb	< 0.01	975/-	965/5
4	11	polblogs	0.03	23843/7272	22821/6975
4	82	foodweb	0.02	71/-	65/2
4	84	celegansneural	< 0.01	759/-	729/11
4	118	polblogs	0.01	1415/498	1307/461
4	131	celegansneural	0.01	112/-	105/3

alone. Lack of other motifs suggests transitivity is the dominant and only force underlying the relations: if two nodes share neighbors, they are likely to be neighbors as well.

In contrast to co-authorship networks, political books, PTT, and protein-protein interaction networks possess richer local structures. Political books and the core set of human PPI networks are enriched with triangles, 3-stars (4 nodes, motif 0) and diamonds (4 nodes, motif 3). PTT and the full set of human PPI networks are enriched with triangles and diamonds. The mechanisms underlying these motifs in distinct networks are certainly different. PPI networks have rich local structures as a variety of possible interaction configurations exist in protein complexes: complex members can bind to each other (triangles), attach to a central member (3-stars), or form a loop (diamonds). Transitivity probably applies in co-purchasing behaviors and PTT user interactions, and co-purchasing behavios may also have the tendency to follow key books (3-stars). It is more difficult to interpret the enrichment of diamonds in political books and PTT networks. One possibility is that diamonds are transient. As time proceeds missing interactions between nodes may be filled.

The network of character co-appearance is enriched with triangles and 3-stars. This observation is sensible as characters often appear in a batch in each chapter (so triangles are enriched), and the central figures appear in multiple chapters (so 3-stars are enriched).

The network of word adjacency is enriched with only 3-stars. Linear orders of words in sentences destroy transitivity and prevent formations of complex motifs. For instance, “good job” and “good mom” are valid phrases while “mom job” or “job mom” are not. In contrast, 3-stars are naturally derived from word adjacency as most adjectives can modify multiple nouns and vice versa.

Two small networks – US football tournaments and the karate club – exhibit distinct characteristics from other networks. The football network is enriched with four network motifs (triangles, 3-stars, diamonds, and triangles with a dangling edge). Knowledge about the rules for arranging tournaments is required in order to interpret the over-representation of these local structures. The karate network is enriched with only diamonds and not with triangles. It consists of two dividing factions centered around the key figures [26]. Shortage of interactions between factions and interactions channeled through key figures may prevent over-representation of triangles.

The four directed networks are categorized into two classes. The food web network is enriched with simply connected, unidirectional structures. Instances include two convergent links (3 nodes, motif 0), unidirectional chains (3 nodes, motif 1; 4 nodes, motif 12), two divergent links (3 nodes, motif 3), diamonds with one source, one sink and two intermediate nodes (4 nodes, motif 81), and diamonds with two sources and two sinks (4 nodes, motif 70). These network motifs are consistent with the nature of a food web: a unidirectional network with clearly defined trophic levels. The remaining three networks – hyperlinks of web documents and US political weblogs and neuronal connections – are enriched with more complex network motifs such as variations of feed-forward loops (3 nodes, motifs 6, 7, 9, 10, 11, 12; 4 nodes, motifs 81, 71, 88, 94, 139, 84) and feed-back loops (3 nodes, motifs 9 and 8; 4 nodes, motifs 77). Variations of feed-forward loops are prominent in both hyperlinks and neuronal connections. Similar to gene regulatory networks, feed-forward loops may confer selective advantages in neuronal networks thus are over-represented. Furthermore, feed-forward loops are enriched in hyperlinks of web documents or political weblogs perhaps due to the existence of multiple paths connecting two documents. In contrast, only three variations of feed-back loops (cycles) are enriched in directed networks. Feed-back loops are prevalent in biological networks including neuronal networks. However, most feed-back loops in neuronal networks likely consist of more than 2 levels. Thus short feed-back loops such as motif 8 are rare.

Absence of network motif enrichments also provides important insight about the nature of interactions. Chains (3 nodes, motif 0; 4 nodes, motif 1) are not significantly enriched in any undirected network despite their overwhelming occurrences. 3-chains are more abundant than triangles in each undirected network, yet perturbed networks often contain even more 3-chains than empirical networks. Under-representation of chains in empirical networks is likely attributed to the prevalence of transitivity as edges connecting common neighbors create triangles and destroy 3-chains.

Unlike triangles, 4-node cliques are not enriched in any undirected network. In principle, transitivity should extend beyond 3 nodes and facilitate the formation of higher-order cliques. Yet this tendency is counter-balanced by the difficulty of maintaining all links in large cliques. Consequently, 4-node cliques are rare and under-represented in empirical networks.

In directed networks, motifs compatible with 3-chains are enriched only in the food web. These enriched motifs consist of only unidirectional links (motifs 0, 1, 3). In contrast, chain motifs possessing bidirectional links are not enriched in any network. Under-representation of these chain motifs suggests that chains emerge from networks with clearly defined levels such as food webs.

We only consider 3-node and 4-node motifs in this study as counting of higher-order motifs and generating valid edge swaps are much more time consuming. However, higher-order network motifs may confer functional/mechanistic importance and be enriched in some networks. More efficient algorithms for counting and edge swapping are required in order to investigate the distributions of higher-order motifs.

The observations of network motif enrichment are purely phenomenological. Certain local structures are over-represented in some networks, yet the causes or mechanisms driving the over-representation of network motifs remain unknown. In-depth studies of specific instances of network motifs and addition information of the interactions are needed in order to understand the underlying mechanisms.

ACKNOWLEDGEMENTS

We thank the inputs from Jing-Shiang Hwang and Hwai-Chung Ho during the preparation of the manuscript. The study is supported by the frontier project grant of Academia Sinica, grant number AS-100-TP2-C01.

REFERENCES

[1] Barabasi, A.L., Albert, R.: Power-law distribution of the world wide web. Science 287 (5461):2115, 2000.

[2] Jeong H., Mason, S.P., Barabasi, A.L.: Lethality and centrality in protein networks. Nature 411:41-42, 2001.

[3] Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature 393, 440-442, 1998.

[4] Girvan M., Newmann, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99(12):7821-7826, 2002.

[5] Bickel, P.J., Chen, A.: A nonparametric view of network models and Newman-Girvan and other modularities. Proc. Natl. Acad. Sci. USA 106(50):21068-21073, 2009.

[6] Expert P., Evans T.S., Blondel V.D., Lambiotte R.: Uncovering space-independent communities in spatial networks. Proc. Natl. Acad. Sci. USA 108(19):7663-7668, 2011.

[7] Axelrod, R.: The dissemination of culture. J. Conflict Resolution 41(2):203-226, 1997.

[8] May , R.M.: Will a large complex system be stable? Nature 238:413-414, 1972.

[9] Kearns, M., Surim S., Montfort, N.: An experimental study of the coloring problem on human subject networks. Scinece 313:824-827, 2006.

[10] Friedkin, N.E.: Norm formation in social influence networks. Social Networks 23:167-189, 2001.

[11] Fowler, J.H., Settle, J.E., Christakis, N.A.: Correlated genotypes in friendship networks. Proc. Natl. Acad. Sci. USA 108(5):1993-1997, 2011.

[12] Liu, F., Lee, H.J.: Use of social network information to enhance collaborativefiltering performance. Expert Systems with Applications 37(7):4772-4778, 2010.

[13] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science 298: 824-827, 2002.

[14] Mangan S., Alon, U.: Structure and function of the feed-forward loop network motif. Proc. Natl. Acad. Sci. USA 100(21):11980-11985, 2003.

[15] Kalir, S., Mangan, S., Alon, U.: A coherent feed-forward loop with a SUM input function protects flagella production in *Escherichia coli*. Mol. Syst. Biol., msb41000010:E1-E6, 2005.

[16] Granovetter, M.S.: The strength of weak ties. American J. Socilogy 78:1360-1380, 1973.

[17] Burt, R.S., Kenz, M.: Kinds of third-party effects on trust. Rationality and Society 7(3):255-292m 1995.

[18] Faust, K.: Very local structure in social networks. Sociological Methodology, 32:209-256, 2007.

[19] Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation networks of *Escherichia coli*. Nat. Genet., 31:64-82, 2002.

[20] Alon, U.: An introduction to systems biology – design principles of biological circuits. Chapman & Hall/CRC, 2007.

[21] Albert, R., Jeong, H., Barabasi, A.L.: Diameter of the world wide web. Nature 401:130-131, 1999.

[22] Adamic, L.A., Glance, N.: The political blogosphere and the 2004 US Election. In Proc. WWW-2005 Workshop on WEblogging Ecosystem, 2005.

[23] Rossberg, A.G., Matsuda, H., Amemiya, T., Itoh, K.: Food webs: experts consuming familues of experts. J. Theor. Biol. 241:552-563, 2006.

[24] Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA 98:404-409, 2001.

[25] Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104, 2006.

[26] Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropological Res. 33:452-473, 1977.

[27] Knuth, D.E.: The Stanford GraphBase: A Platform for Combinatorial Computing. Addison-Wesley, Reading, MA, 1993.

[28] Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., and Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology 54, 396-405, 2003.

[29] http://www.orgnet.com/.

[30] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 32 (Database Issue): D449-451, 2004.

[31] PTT Bulletin Board System, telnet://ptt.cc.