MOTIFS ANALYSIS

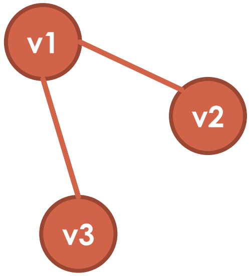
Bioinformatics
ay 2018-2019

Manuela Petti

12/12/2018

Graphs

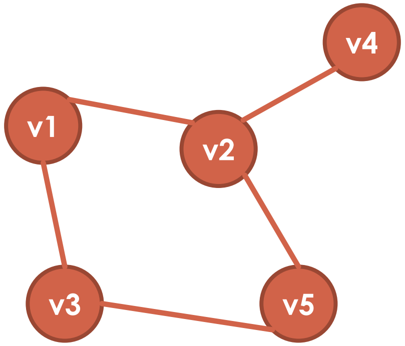
- Graph *G*=(*V*,*E*) is a set of vertices *V* and edges *E*
  - *V = {v1, v2, v3, v4, v5}*
  - *E = {(v1, v2), (v1, v3), (v2, v4), (v2, v5) , (v3, v5)}*



- A subgraph *G'* of *G* is induced by some $V' \subset V$ and $E' \subset E$

For example, *V' = {v1, v2, v3}* and *E' = {(v1, v2), (v1, v3)}*

- Graph properties:
  - Directed vs. undirected
  - Weighted vs. unweighted
  - Cyclic vs. acyclic
  - Graph indices

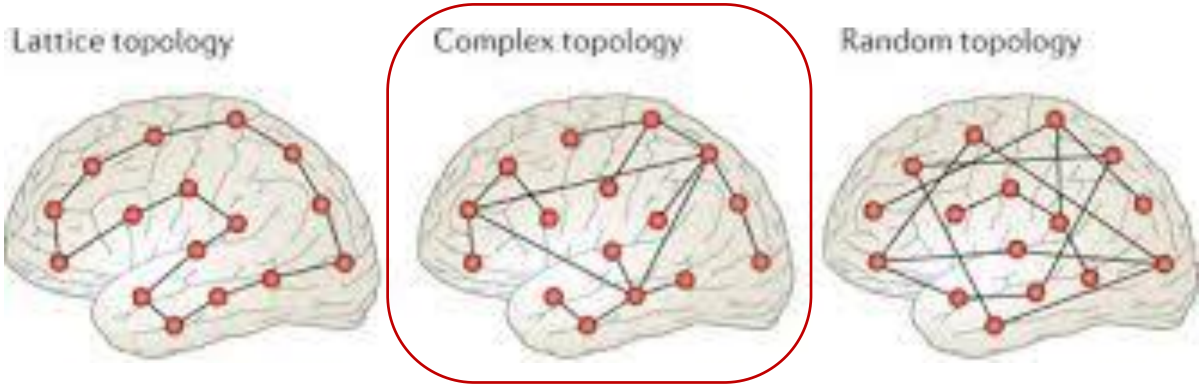


Network in Neuroscience

Small-world networks: rapid integration of information from local, specialized brain areas even when they are distant (Sporns O & Zwi JD, 2004)

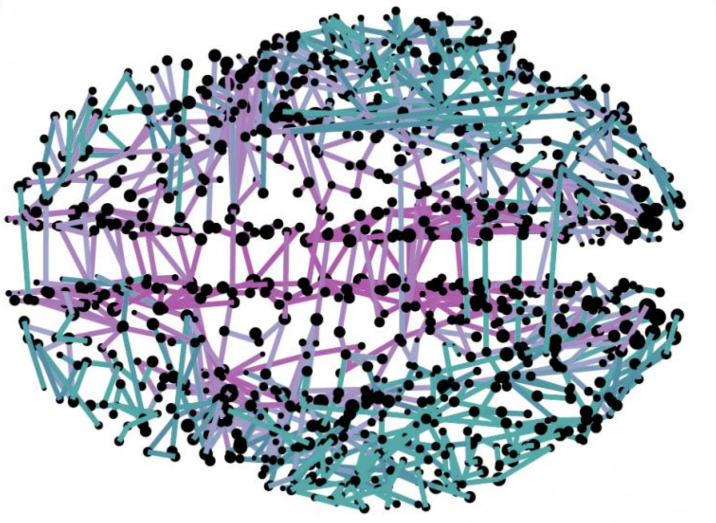
Small-world organizations in human brain networks have primarily been studied in resting state, either using fMRI or M/EEG

- Abnormalities in these resting state networks appear to be relate to neurological and psychiatric diseases
- Changes in small-worldness have been found as effect of aging, in different sleep stages
- Changes in graph measures when subjects performed foot movements and finger tapping



Small-world organizations

Graph theory in Neuroscience



Microscopic

Degree (in-out)

Betweenness centrality

Closeness centrality

eigenvector centrality

Mesoscopic

Motifs

Modularity

Macroscopic

Average Path length

Clustering coefficient

Global efficiency

Local efficiency

go beyond global and local features for understanding the basic structural elements particular to each class of networks

What is a motif?

In a sequence, a motif is a recurring subsequence, a pattern, that is conjectured to have some functional significance

In a network, a motif is a recurring subnetwork conjectured to have some significance, that appears with a higher frequency than it would be expected in “similar” random networks

The concept has been applied to networks of different nature:

- protein-protein interaction (PPI)
- gene transcriptional regulation
- food webs
- brain
- electronic circuits
- software

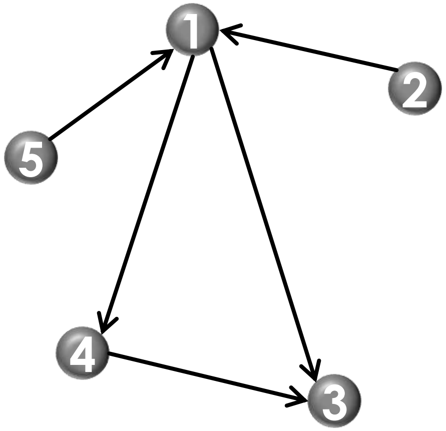
Motif Definition

Network Motifs: Simple Building Blocks of Complex Networks

Patterns of inter-connections
occurring in complex networks at numbers that are significantly higher than those in randomized networks

(Milo et al, Science, 2002)

They encode basic interconnection properties
of complex networks in nature

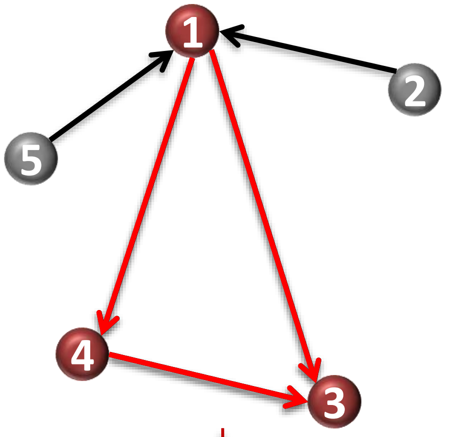


Counting how many times a motif appears in a given network yields a frequency spectrum that contains important information on the network basic building blocks

Motif Definition in Brain Network

STRUCTURAL MOTIFS

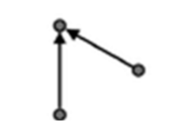
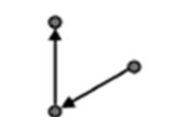
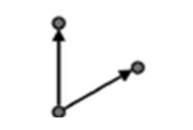
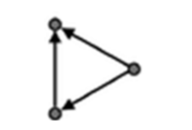
A structural motif of size M is comprised of a specific set of M vertices that are linked by edges.



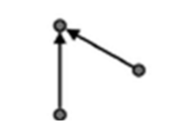
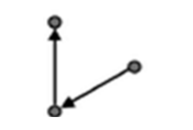
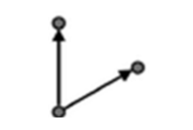
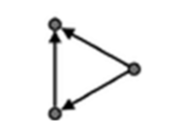
FUNCTIONAL MOTIFS

Functional motifs form a set of sub-graphs of a given structural motif. All such functional motifs consist of the original M vertices of the structural motif to which they belong, but contain only a subset of its edges

STRUCTURAL MOTIFS

			
5→1←2	5→1→3		4←1→3←4
	5→1→4		
	2→1→3		
	2→1→4		

FUNCTIONAL MOTIFS

			
5→1←2	5→1→3	4←1→3	4←1→3←4
1→3←4	5→1→4		
	2→1→3		
	2→1→4		
	1→4→3		

Sporns et al, 2004

Motif Definition

An induced subgraph G_K of a graph G is called a network motif when for a given set of parameters {*p,U,D,N*} and a random ensemble of N similar networks:

1. $Prob(f_{random}(G_K) > f_{original}(G_K)) \leq P$
   (Over-representation)
2. $f_{original}(G_K) \geq U$
   (Minimum frequency)
3. $f_{original}(G_K) - f_{random}(G_K) > D \times f_{random}(G_K)$
   (Minimum deviation)

$$Z = \frac{f_{original}(G_k) - \langle f_{random}(G_k) \rangle}{\sigma_{rand}}$$

Where:

- $f_{original}$ is the frequency in the original network
- f_{random} is the frequency in a random network
- *U* is an uniqueness threshold
- *D* is the proportional threshold that ensures the minimum difference between $f_{original}$ and f_{random}

Milo et al (Science, 2002) used {0.01, 4, 0.1, 1000} as the set of parameters

Motif Definition

Anti-motifs are significantly under-represented subnetworks and may also be meaningful

They are subgraphs that satisfy the following:

(i) the probability that they appear in randomized networks fewer times than in the real network is lower than p

(ii) $f_{random}(G_K) - f_{original}(G_K) > D \times f_{random}(G_K)$

Possible interpretation

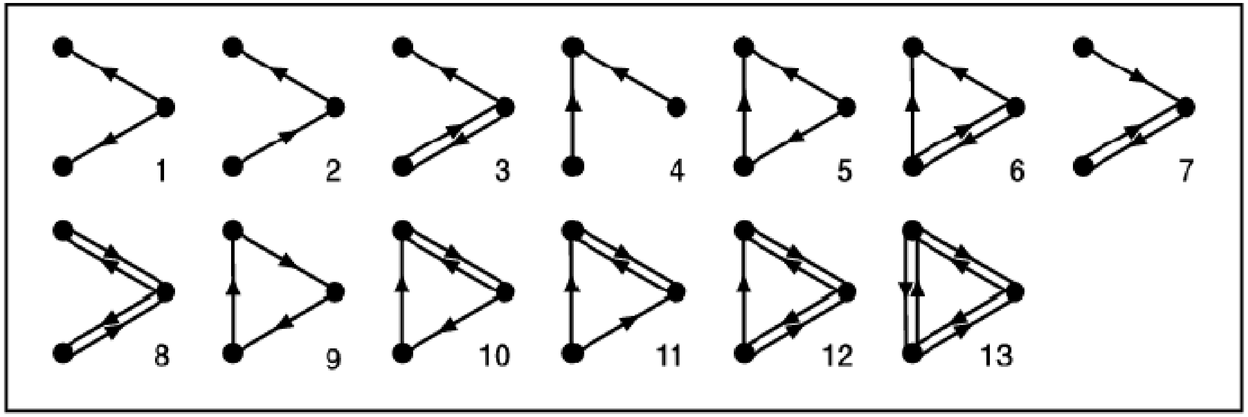
- the motifs appeared because of constraints in the way the network was developed, thus being related to the evolution of the whole complex system.

- the motifs appeared in relation to the classes of networks based on types of motifs found

What rules underlie the organization of the particular types of networks that we see in complex brains?
It is likely that, as networks become more complex, already existing simpler networks are largely preserved, extended, and combined, while it is less likely that complex structures are generated entirely de novo.

3-node motifs

- 13 different isomorphic types of 3-node connected subgraph



- There are:

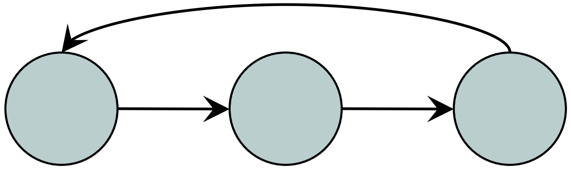
199 4-node subgraphs

9364 5-node subgraphs

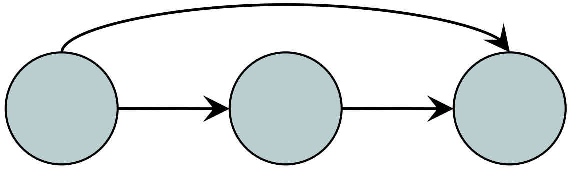
……

Network Motifs

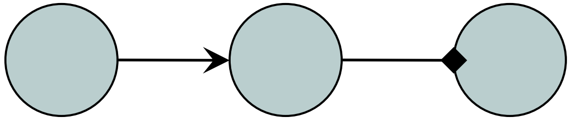
Feedback



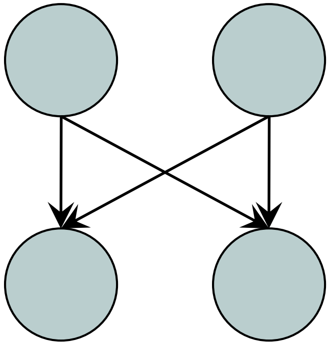
feed-forward



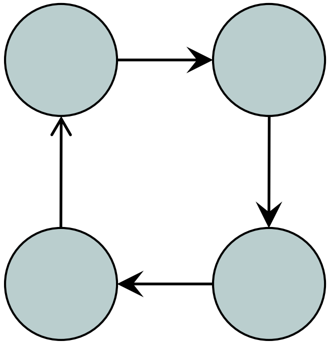
Three chain



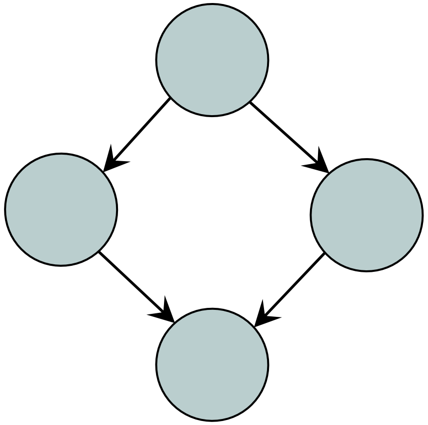
Bi-fan



4-node feedback



Bi-parallel



Network Motifs Detection

Finding the motifs is a computationally hard task

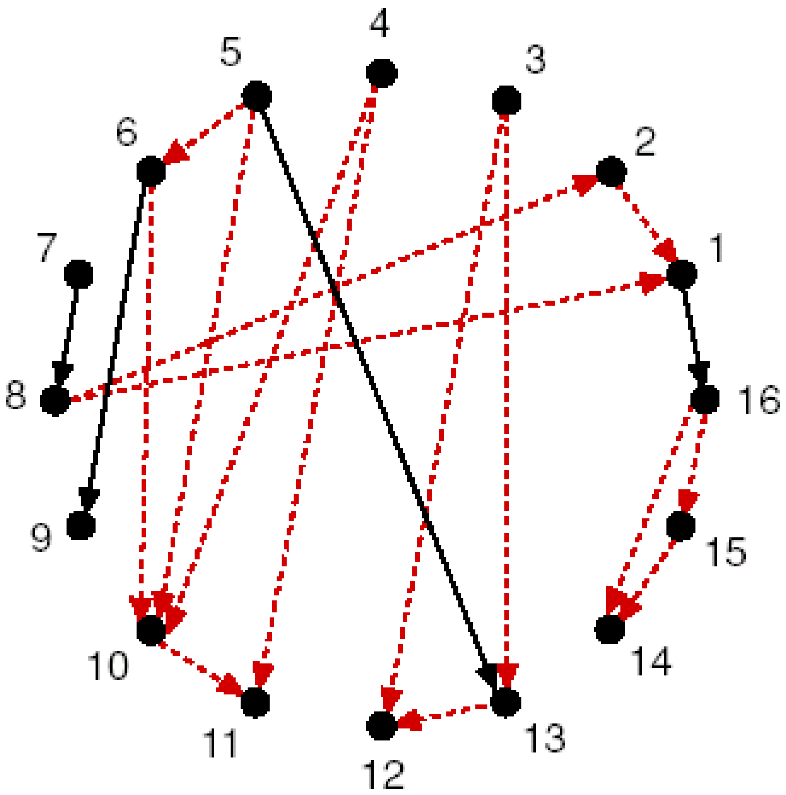
As the size of the motifs increases, the
time needed to calculate them grows exponentially.

Hence, an exhaustive computation of all motifs of a network is typically reduced to very small sizes in order to obtain results in a reasonable amount of time

To guarantee that when searching for k-motifs, the frequency of (k−1)-motifs would be the same, ensuring that the significance of a particular pattern does not simply derive from its subpatterns

Network Motifs Detection

- Find n-node subgraphs in real graph.
- Find all n-node subgraphs in a set of proper random graphs
- Assign Z-score for each subgraph.
- Subgraphs with high Z-scores are denoted as **Network Motifs**.

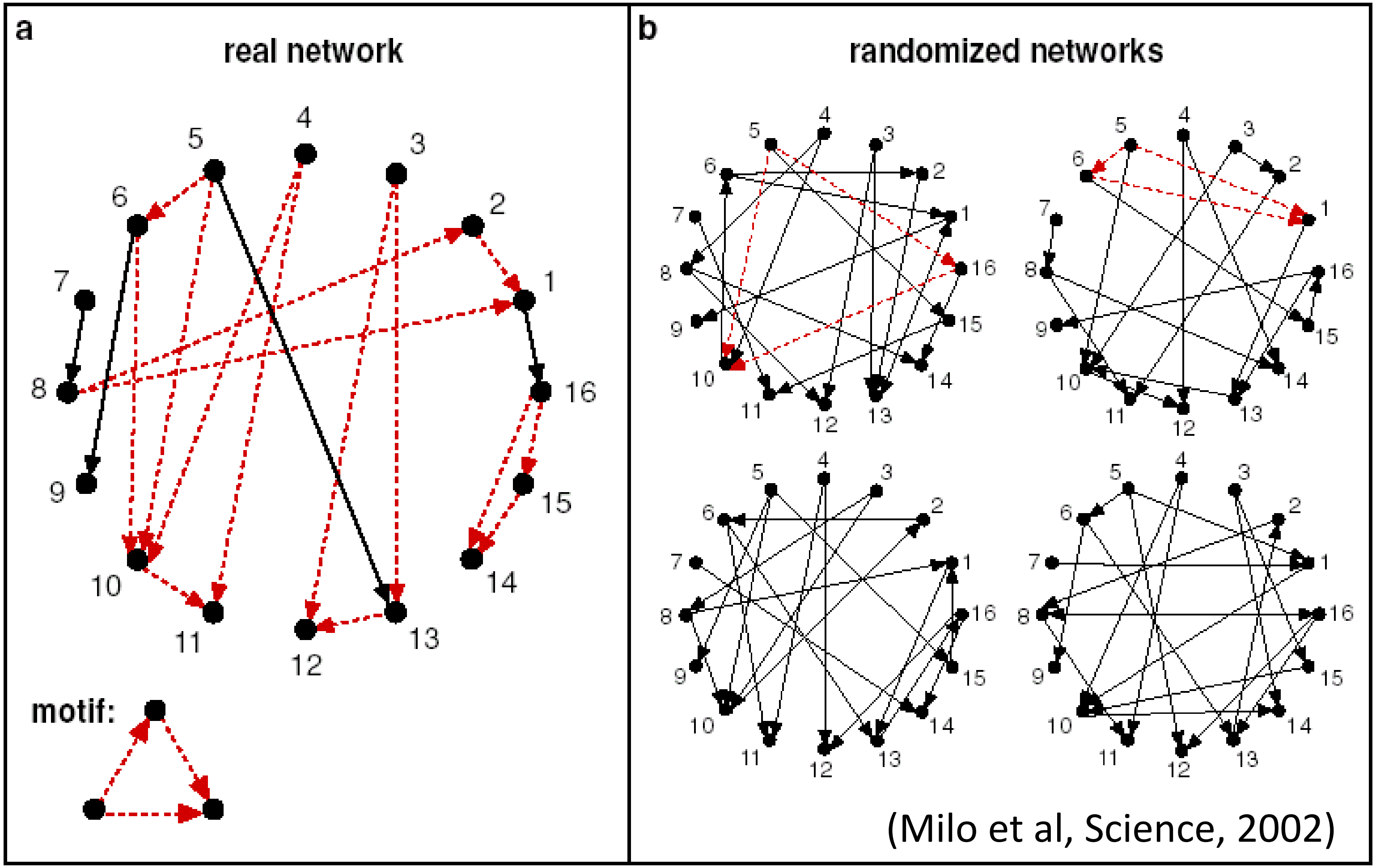


$$Z = \frac{f_{original}(G_k) - \langle f_{random}(G_k) \rangle}{\sigma_{rand}}$$

Network Motifs Detection

Network motifs are patterns that recur much more frequently in the real network than in an ensemble of randomized networks.

Each node in the randomized networks has the same number of incoming and outgoing edges as does the corresponding node in the real network.



(Milo et al, Science, 2002)

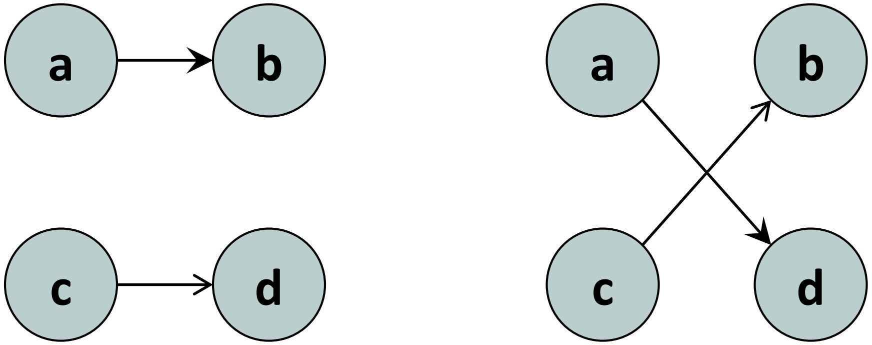
Network Motifs Detection

Main tasks in detecting network motifs:

1) generating the set of proper random networks
2) counting the subgraphs in the real network and in random networks.

Generation of random graphs

- Algorithm A
  - Employ a Markov-chain algorithm based on starting with the real network and repeatedly swapping randomly chosen pairs of connections (a→b, c→d is replaced by a→d, c→b) until the network is well randomized.
  - Switching is prohibited if connections a→d or c→b already exist

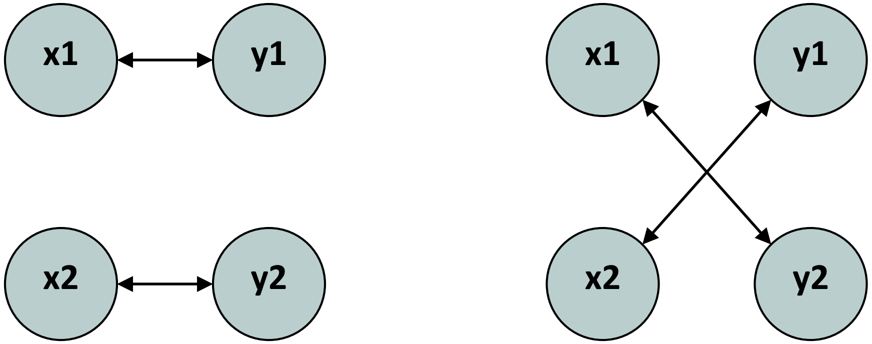


(Milo et al, Science, 2002)

Generation of random graphs

- Algorithm A
  - It allows to preserve the numbers of the in- and outgoing edges for each node, as well as the number of mutual edges (X ↔ Y) for each node
  - A double edge is switched only with a different double edge

(X1 ↔ Y1, X2 ↔ Y2 to X1 ↔ Y2, X2 ↔ Y1)



(Milo et al, Science, 2002)

Generation of random graphs

- Algorithm B
  - Each network was presented as a connectivity matrix M: M_{ij} = 1 if there is a connection directed from node i to node j, and 0 otherwise
  - The goal is to create a randomized connectivity matrix M_{rand}, which has the same number of nonzero elements in each row and column as the corresponding row and column of the real connectivity matrix

(Milo et al, Science, 2002)

Motifs discovery algorithms

- ➢ Enumeration algorithms

  To look for motifs of size k, first enumerate all k-subgraphs of the original graph and then calculate a subgraph census

  A subgraph census is performed on each random network →the statistical significance of the motifs on the original network is evaluated and the ones over-represented are reported

PROBLEM

with the complete census the
number of existing subgraphs
grows exponentially increasing the
size of the network or the size of
the subgraphs themselves

Run time complexity

- The performance of this algorithm scales with the total number of *n*-node subgraphs in the network.
- The number of subgraphs and the algorithm runtime also increase dramatically for subgraphs with $n \geq 5$.

to sacrifice accuracy using a probabilistic approximation algorithm

Motifs discovery algorithms

- Approximate algorithms

  Sampling of a determined number of k-subgraphs on the original and on the random networks. We can then use their concentration to obtain an approximated z-score and therefore calculate an approximate significance (Mfinder, FanMod, Rand-ESU)

Finding *n*-node subgraphs for *n* ≥5 is much easier

Approximate Algorithms

MFINDER: it picks at random edges of the input graph until a set of k nodes obtained to get sample sub-graph and assigns weights to the samples to correct the non-uniform sampling.

It scale well with large networks, but does not scale well with large motifs.

Subgraph Concentrations

- Let N_i be the number of appearances of subgraphs of type *I*

- The concentration of *n*-node subgraphs of type *i* is the ratio between their number of appearances and the total number of *n*-node connected subgraphs in the network:

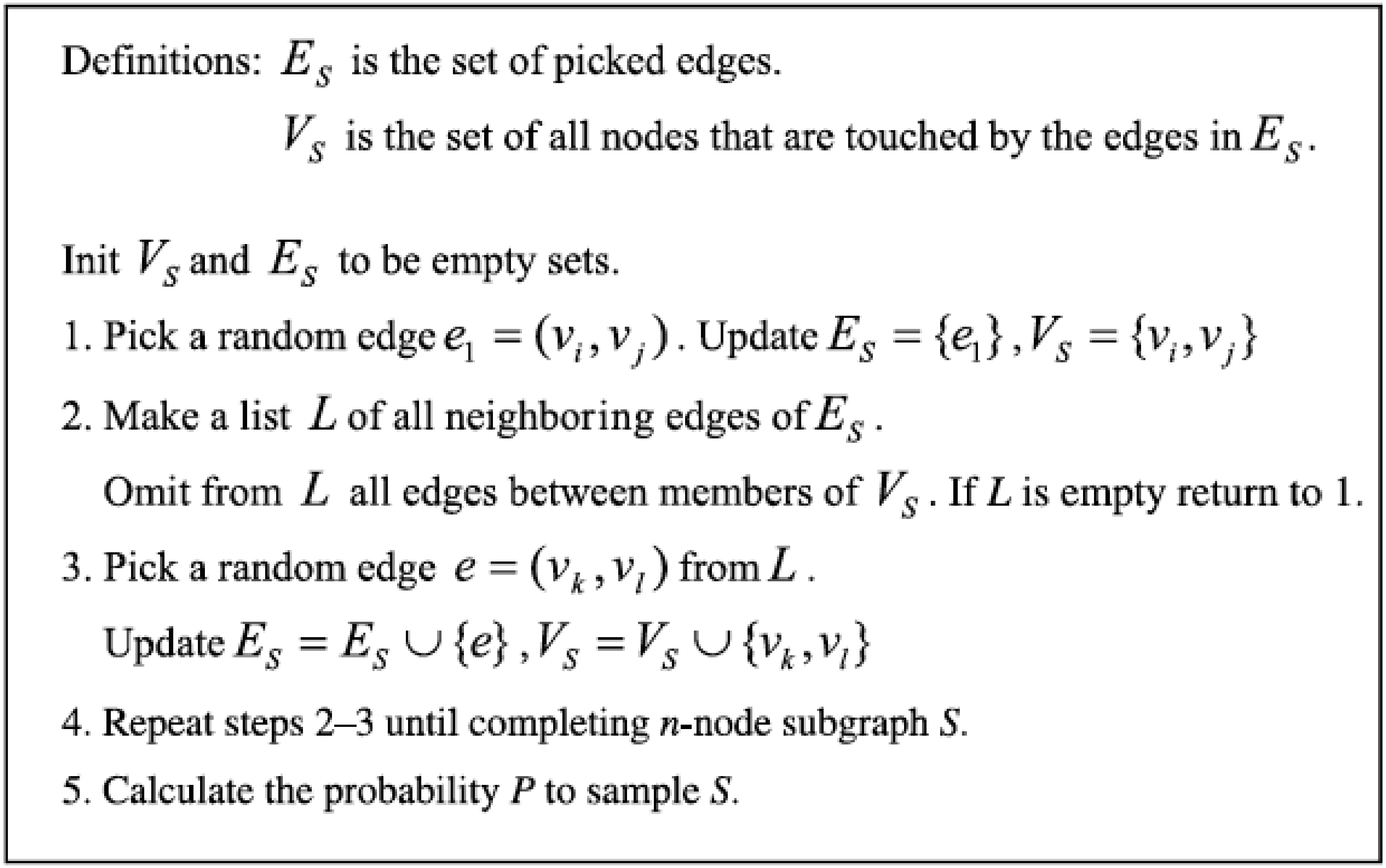
$$C_i = \frac{N_i}{\sum_i N_i}$$

Subgraphs Sampling

The algorithm samples n-node subgraphs by picking random connected edges until a set of n nodes is reached

- pick a random edge from the network and then expand the subgraph iteratively by picking random neighboring edges until the subgraph reaches n nodes
- for each random choice of an edge, in order to pick an edge that will expand the subgraph size by one, prepare a list of all such candidate edges and then randomly choose an edge from the list
- the sampled subgraph is defined by the set of n nodes and all the edges that connect between these nodes in the original network

Subgraphs Sampling



```
Definitions: $E_S$ is the set of picked edges.
             $V_S$ is the set of all nodes that are touched by the edges in $E_S$.

Init $V_S$ and $E_S$ to be empty sets.
1. Pick a random edge $e_1 = (v_i, v_j)$. Update $E_S = \{e_1\}, V_S = \{v_i, v_j\}$
2. Make a list $L$ of all neighboring edges of $E_S$.
   Omit from $L$ all edges between members of $V_S$. If $L$ is empty return to 1.
3. Pick a random edge $e = (v_k, v_l)$ from $L$.
   Update $E_S = E_S \cup \{e\}, V_S = V_S \cup \{v_k, v_l\}$
4. Repeat steps 2–3 until completing $n$-node subgraph $S$.
5. Calculate the probability $P$ to sample $S$.
```

Sampling Probability

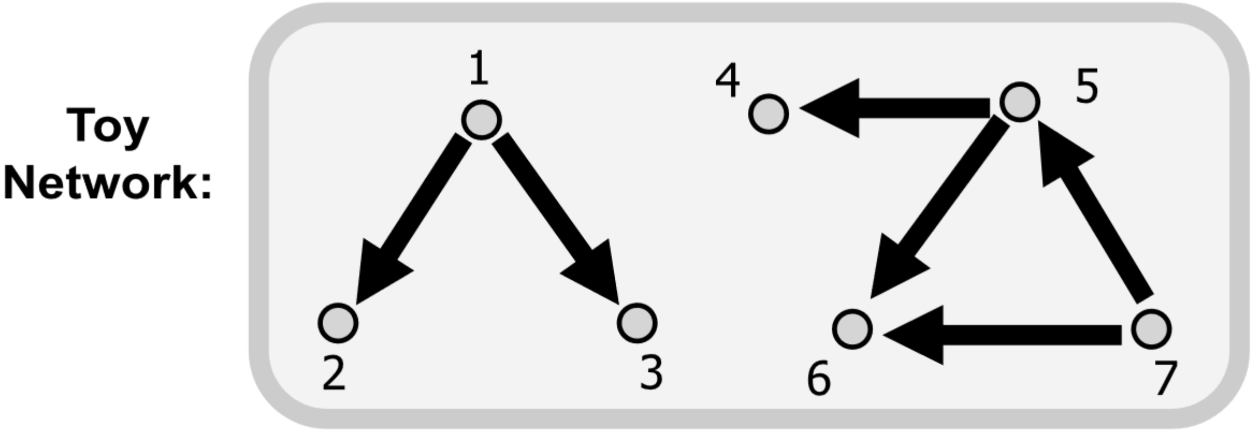
To sample an n-node subgraph, an ordered set of n-1 edges is iteratively randomly picked.
In order to compute the probability, P, of sampling the subgraph, we need to check all such possible ordered sets of n-1 edges [denoted as (n-1)-permutations] that could lead to sampling of the subgraph

The probability of sampling the subgraph is the sum of the probabilities of all such possible ordered sets of n-1 edges:

$$P = \sum_{\sigma \in S_m} \prod_{E_j \in \sigma} \Pr[E_j = e_j | (E_1, \ldots, E_{j-1}) = (e_1, \ldots, e_{j-1})].$$

where S_m is the set of all (n-1)-permutations of the edges from the specific subgraph edges that could lead to a sample of the subgraph. Ej is the j-th edges in a specific (n-1)-permutation

Sampling Probability



Probability to sample {1,2,3}:
There are 2 possibilities to sample {1,2,3}:
1. Pick first (1,2): Pr=1/E=1/6.
then pick (1,3): Pr=1.
Pr[(1,2) then (1,3)]=1/6*1=1/6.
2. Pick first (1,3):Pr=1/E=1/6.
then pick (1,2): Pr=1.
Pr[(1,3) then (1,2)]=1/6*1=1/6.
In Total: Pr[{1,2,3}] = 1/6 + 1/6 = 1/3=12/36

Probability to sample {4,5,6}:
There are 2 possibilities to sample {4,5,6}:
1. Pick first (5,4): Pr=1/E=1/6.
then pick (5,6): Pr=1/2.
Pr[(5,4) then (5,6)]=1/6*1/2=1/12
2. Pick first (5,6): Pr=1/E=1/6.
then pick (5,4): Pr=1/3.
Pr[(5,6) then (5,4)]=1/6*1/3=1/18.
In Total: Pr[{4,5,6}] = 1/12 + 1/18 = 5/36

Fig. 2. Different probabilities of sampling different subgraphs. Example of a toy network with seven nodes and six directed edges. The probabilities of sampling two different three-nodes subgraphs are different, although they both are of the same subgraph type (V-shaped outgoing edges).

(Kashtan et al, Bioinformatics, 2004)

Calculating the concentrations of n-node subgraphs

- Define score S_i for each subgraph of type i
- Inizialize S_i to 0 for all i
- For every sample, add the weighted score W=1/P to the accumulated score S_i of the relevant type I

$$S_i = S_i + W$$

- After S_T samples, assuming we sampled L different subgraph types, calculate the estimated subgraph concentrations:

$$C_i = \frac{S_i}{\sum_{k=1}^{L} S_{k_i}}$$

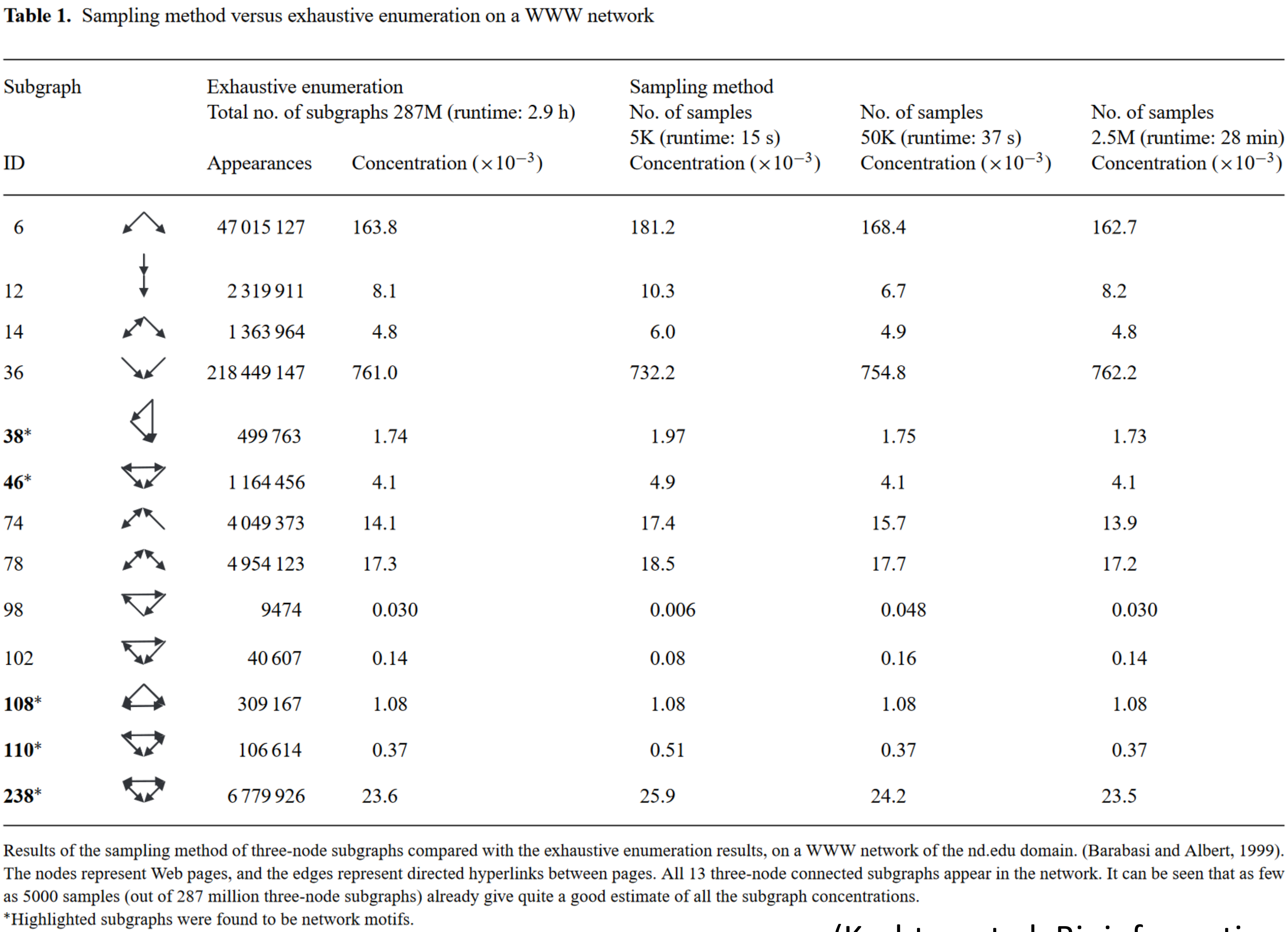


Table 1. Sampling method versus exhaustive enumeration on a WWW network

Subgraph		Exhaustive enumeration Total no. of subgraphs 287M (runtime: 2.9 h)		Sampling method No. of samples 5K (runtime: 15 s)	No. of samples 50K (runtime: 37 s)	No. of samples 2.5M (runtime: 28 min)
ID		Appearances	Concentration ($\times 10^{-3}$)	Concentration ($\times 10^{-3}$)	Concentration ($\times 10^{-3}$)	Concentration ($\times 10^{-3}$)
6		47 015 127	163.8	181.2	168.4	162.7
12		2 319 911	8.1	10.3	6.7	8.2
14		1 363 964	4.8	6.0	4.9	4.8
36		218 449 147	761.0	732.2	754.8	762.2
38*		499 763	1.74	1.97	1.75	1.73
46*		1 164 456	4.1	4.9	4.1	4.1
74		4 049 373	14.1	17.4	15.7	13.9
78		4 954 123	17.3	18.5	17.7	17.2
98		9474	0.030	0.006	0.048	0.030
102		40 607	0.14	0.08	0.16	0.14
108*		309 167	1.08	1.08	1.08	1.08
110*		106 614	0.37	0.51	0.37	0.37
238*		6 779 926	23.6	25.9	24.2	23.5

Results of the sampling method of three-node subgraphs compared with the exhaustive enumeration results, on a WWW network of the nd.edu domain. (Barabasi and Albert, 1999). The nodes represent Web pages, and the edges represent directed hyperlinks between pages. All 13 three-node connected subgraphs appear in the network. It can be seen that as few as 5000 samples (out of 287 million three-node subgraphs) already give quite a good estimate of all the subgraph concentrations.

*Highlighted subgraphs were found to be network motifs.

(Kashtan et al, Bioinformatics, 2004)

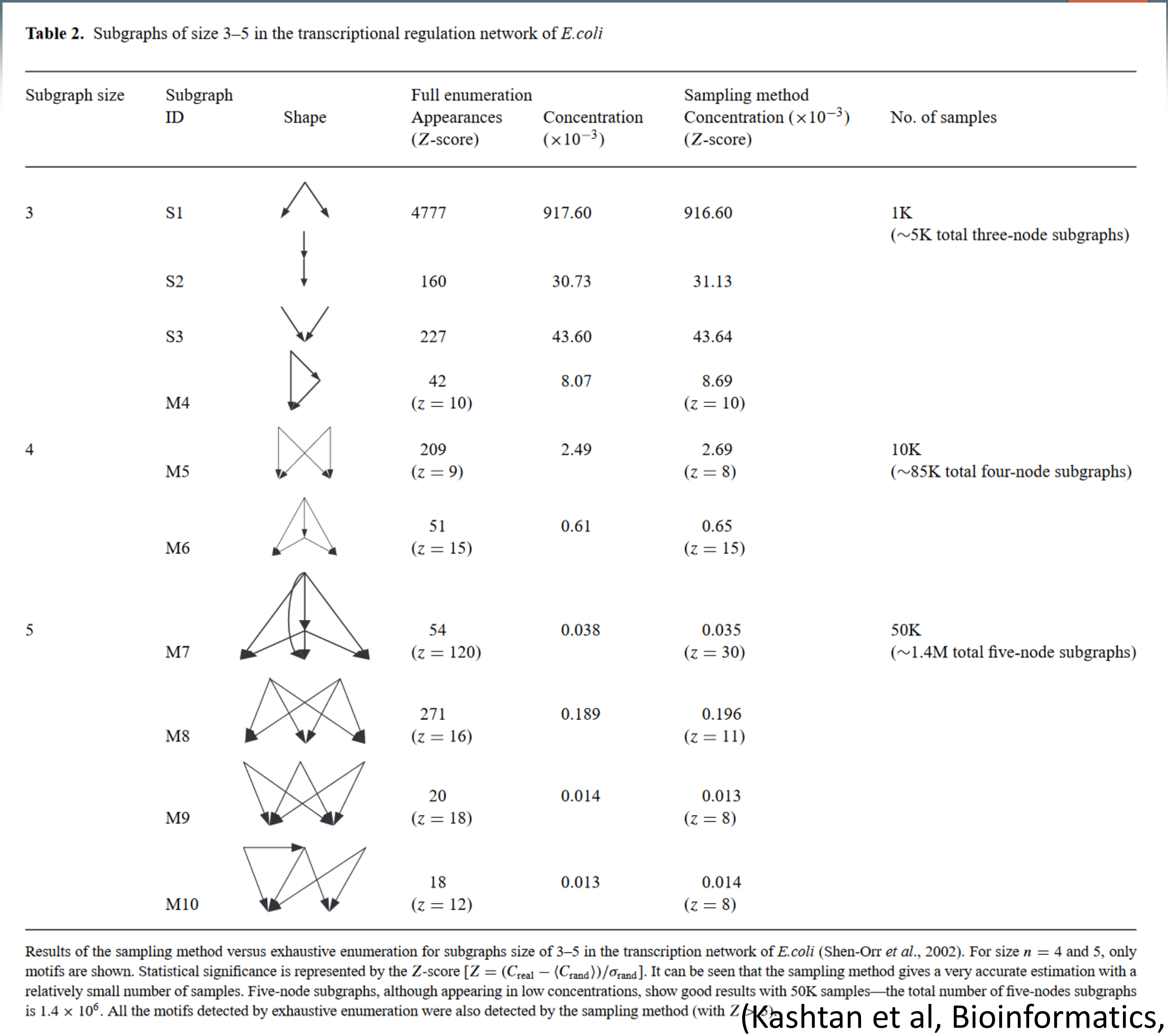


Table 2. Subgraphs of size 3–5 in the transcriptional regulation network of *E.coli*

Subgraph size	Subgraph ID	Shape	Full enumeration		Sampling method	
			Appearances (Z-score)	Concentration ($\times 10^{-3}$)	Concentration ($\times 10^{-3}$) (Z-score)	No. of samples
3	S1		4777	917.60	916.60	1K (~5K total three-node subgraphs)
	S2		160	30.73	31.13	
	S3		227	43.60	43.64	
	M4		42 ($z = 10$)	8.07	8.69 ($z = 10$)	
4	M5		209 ($z = 9$)	2.49	2.69 ($z = 8$)	10K (~85K total four-node subgraphs)
	M6		51 ($z = 15$)	0.61	0.65 ($z = 15$)	
5	M7		54 ($z = 120$)	0.038	0.035 ($z = 30$)	50K (~1.4M total five-node subgraphs)
	M8		271 ($z = 16$)	0.189	0.196 ($z = 11$)	
	M9		20 ($z = 18$)	0.014	0.013 ($z = 8$)	
	M10		18 ($z = 12$)	0.013	0.014 ($z = 8$)	

Results of the sampling method versus exhaustive enumeration for subgraphs size of 3–5 in the transcription network of *E.coli* (Shen-Orr *et al*., 2002). For size $n = 4$ and 5, only motifs are shown. Statistical significance is represented by the Z-score [$Z = (C_{\mathrm{real}} - \langle C_{\mathrm{rand}} \rangle)/\sigma_{\mathrm{rand}}$]. It can be seen that the sampling method gives a very accurate estimation with a relatively small number of samples. Five-node subgraphs, although appearing in low concentrations, show good results with 50K samples—the total number of five-nodes subgraphs is 1.4×10^6. All the motifs detected by exhaustive enumeration were also detected by the sampling method (with $Z > 6$).

(Kashtan et al, Bioinformatics, 2004)

Sampling Method for Subgraph counting

- Kashtan *et al.:* “Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs”; Bioinformatics, 2004.
- This algorithm samples subgraphs in order to estimate their relative frequency.
- The runtime of the algorithm asymptotically does not depend on the network size.
- Surprisingly, few samples are needed to detect network motifs reliably.

http://www.weizmann.ac.il/mcb/UriAlon/research/network-motifs

https://sites.google.com/site/bctnet/Home