Network Motifs

and

Efficient Counting of Graphlets

Yuval Shavitt
School of Electrical Engineering

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

shavitt@eng.tau.ac.il
http://www.netDIMES.org
http://www.eng.tau.ac.il/~shavitt

Network Motifs: Why

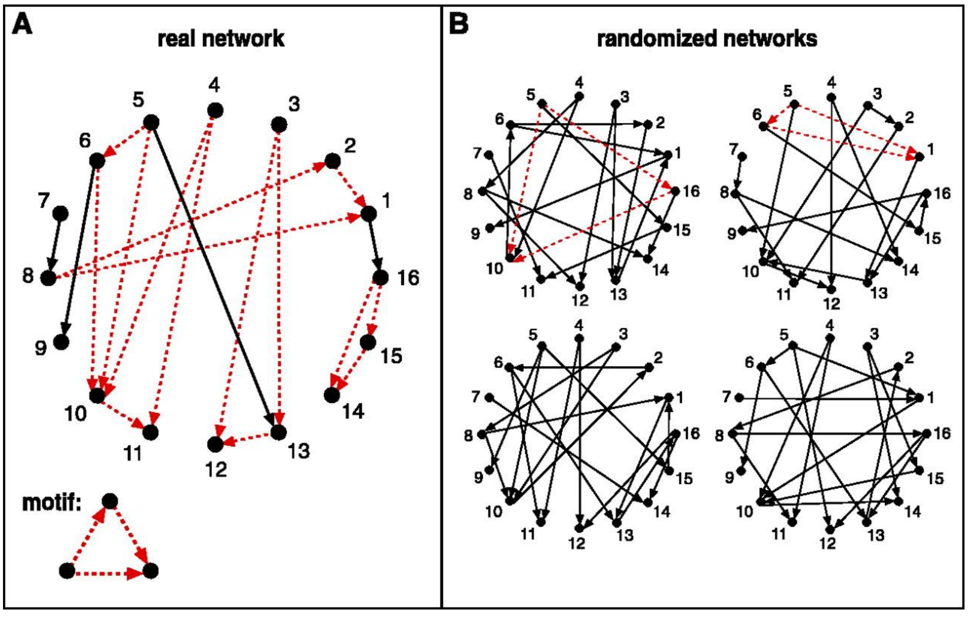
- Complex networks appear in all areas of science
- Share global properties: power-laws, small world, long tail.
- We want to uncover their structural design principles.
  - Use local sub-structure

R. Milo *et al*. Network motifs: Simple building blocks of complex networks. *Science*, 298:824--827, 2002.

Definition

- Network motifs: patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks.

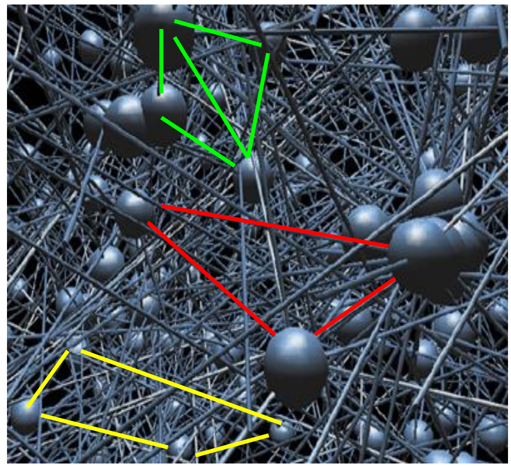
$$Z = (N_{\text{real}} - N_{\text{rand}})/\text{S.D.}$$



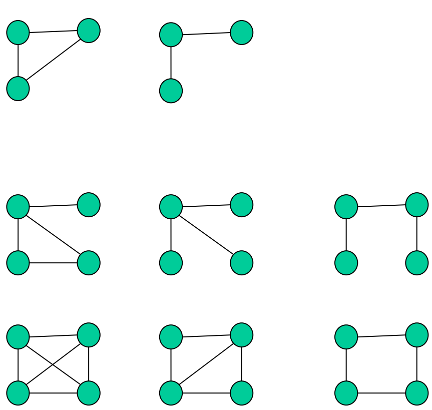
Graphlets

A large complex networks has many small subgraphs:

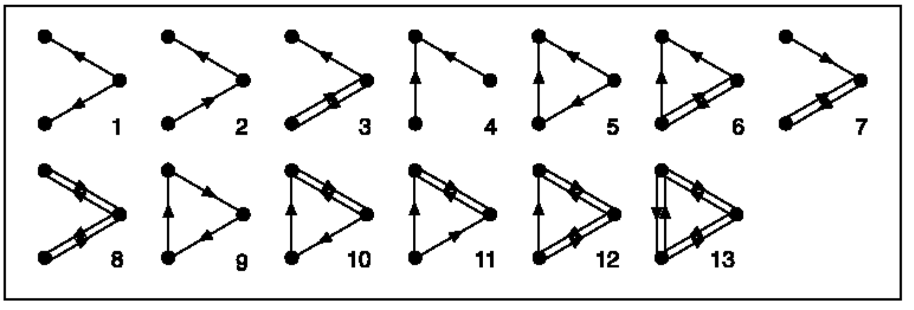
- How many Δ?
- How many □?
- How many ⍂?



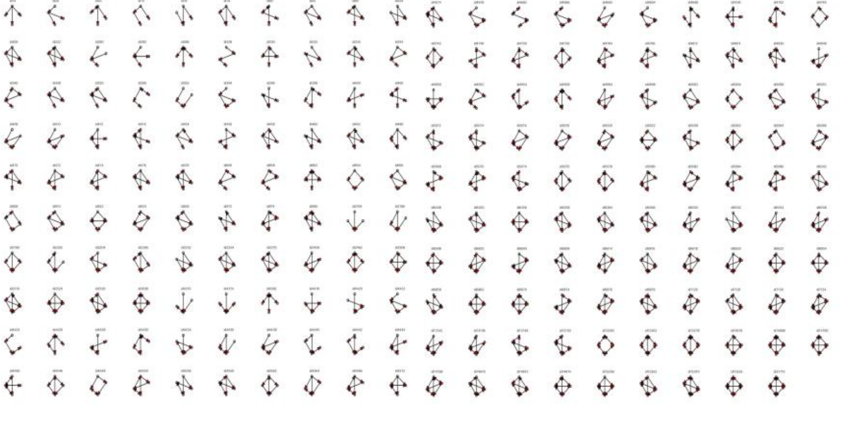
Counting *All* possible Graphlets



All 3-node Directed Graphlets

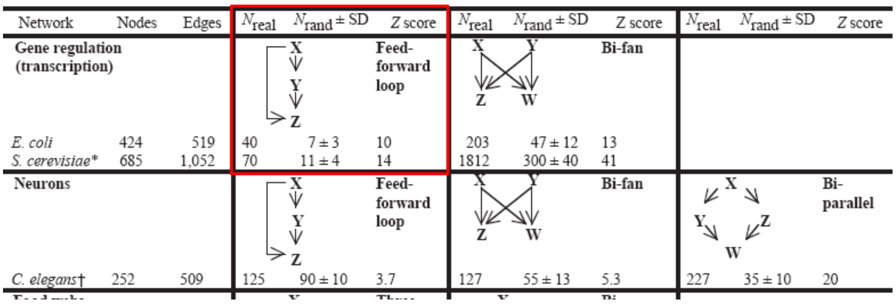


199 4-node directed connected graphlets



5 nodes: 9364
6 nodes: 1,530,843

Motifs in Biological Networks

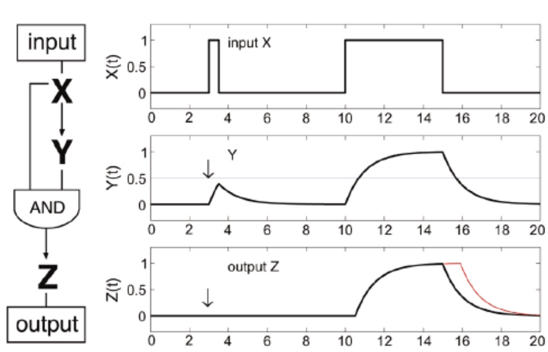


Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)				X → Y → Z, X → Z	**Feed-forward loop**		X, Y → Z, W	**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons				X → Y → Z, X → Z	**Feed-forward loop**		X, Y → Z, W	**Bi-fan**		X → Y, Z → W	**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20

Some motifs are clearly significant

$Z = (N_{real} - N_{rand})/S.D.$

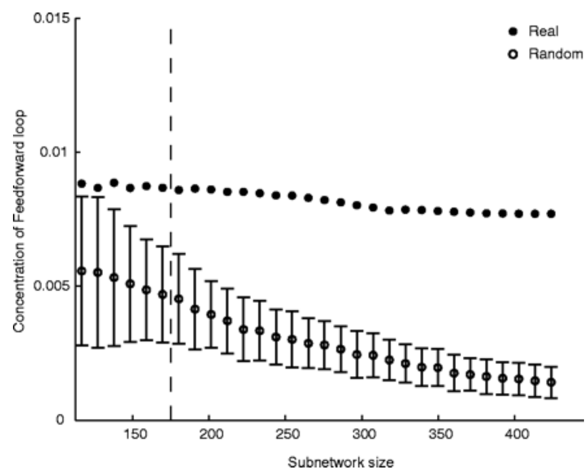
The Gene regulation network of *Escherichia coli*



[Shen-Orr *et al*., Nature Genetics 2002]

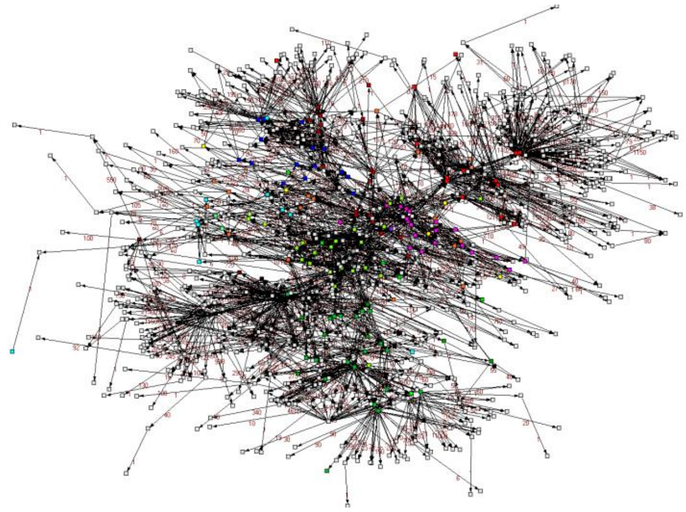
Motif Appearance in Subnetworks

E. coli transcription network

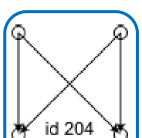
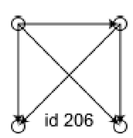
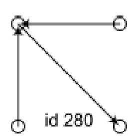
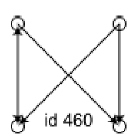
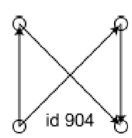


Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

The IP Interface Graph



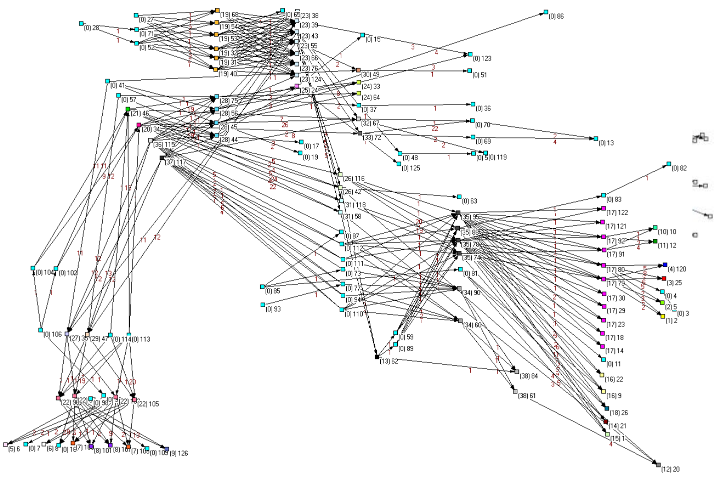
IP Interface Graph

AS Number	Z-Score				
	 id 204	 id 206	 id 280	 id 460	 id 904
AS6395	377	-	9.51	43.84	148.39
AS5111	329.29	36.42	-	74.63	73.57
AS3549	154.8	5.38	37.87	19.51	-

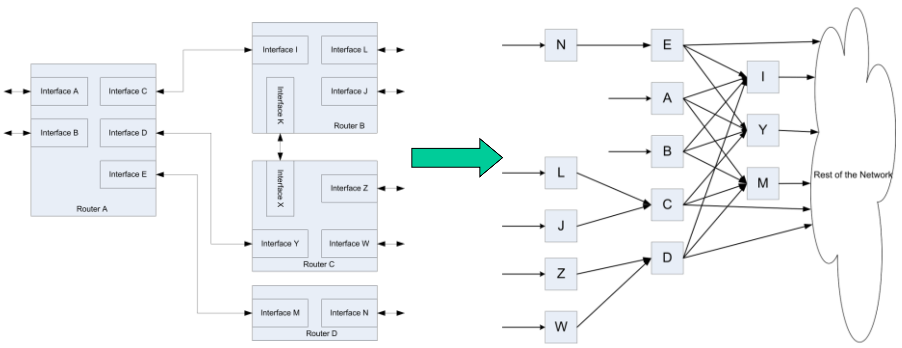
Traceroute create graphs which tend to have many small bi-partite.
WHY?

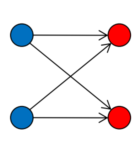
[Feldman & Shavitt, Globecom 08]

The IP Interface Graph

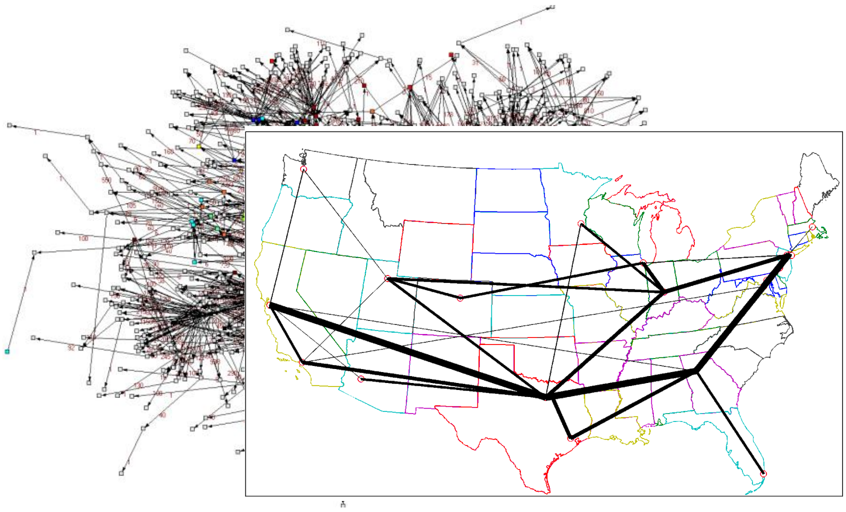


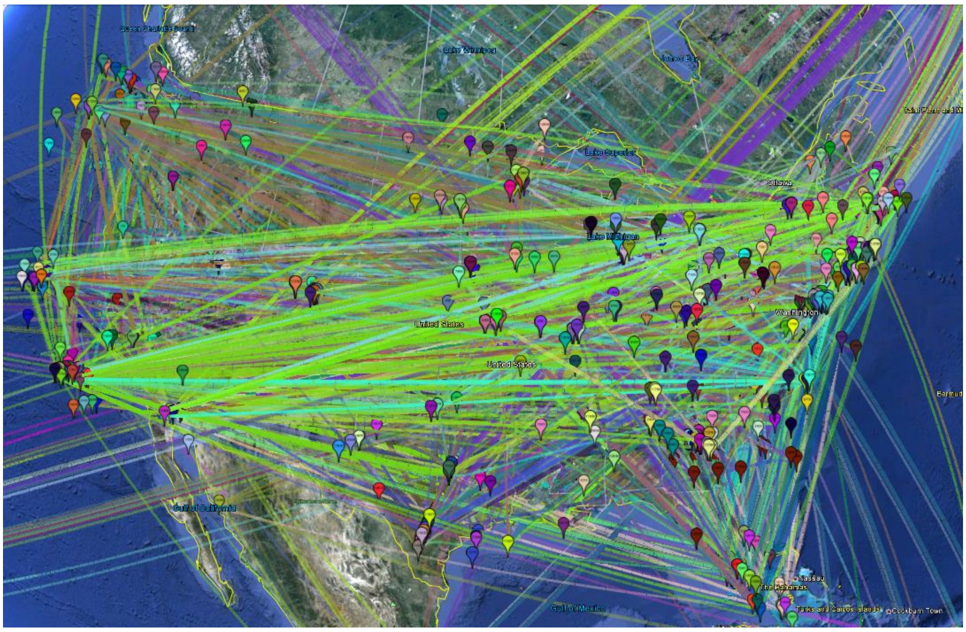
Using Bi-Fan to find Internet PoPs





We use Bi-Fans to Auto-find PoPs





[Feldman, Shavitt, & ZIlberman, *Comp. Net.* Feb. 2012]

Comparing Networks by Local Structure

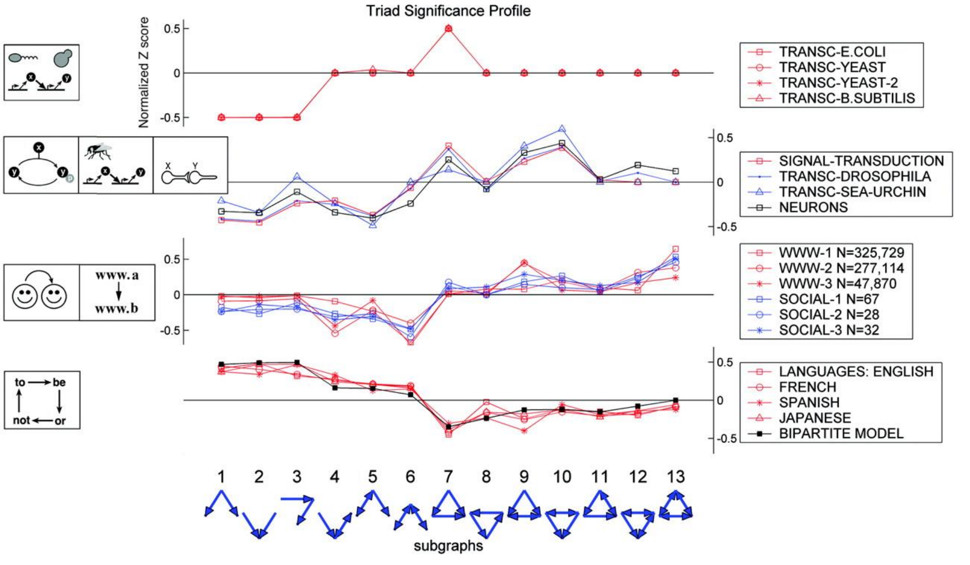
$Z_i = (N_{\mathrm{real}_i} - \langle N_{\mathrm{rand}_i} \rangle)/\mathrm{std}(N_{\mathrm{rand}_i})$

Significance Profile: $SP_i = \frac{Z_i}{\sqrt{\sum_{k=1}^{n} Z_i^{\,2}}}$

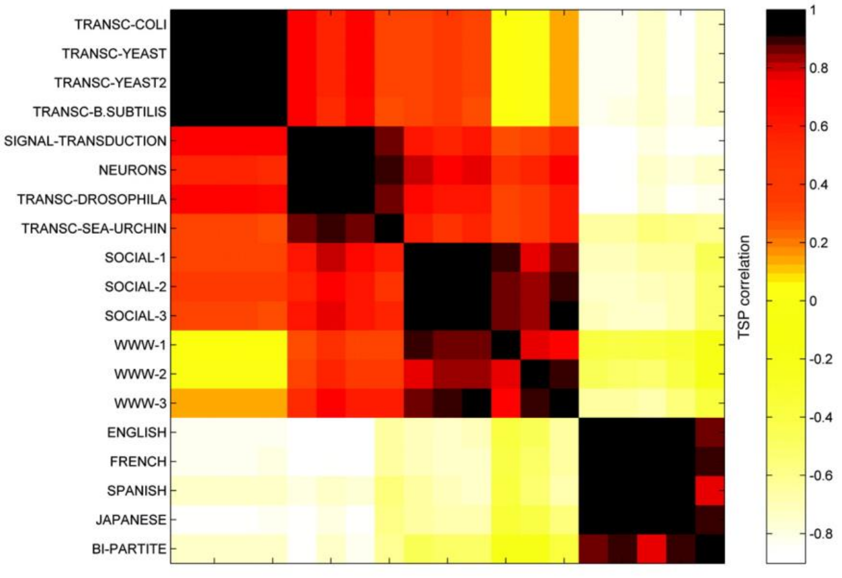
Motifs in large graphs tend to have higher Z scores.

The normalization emphasizes the relative significance of subgraphs, rather than the absolute significance.

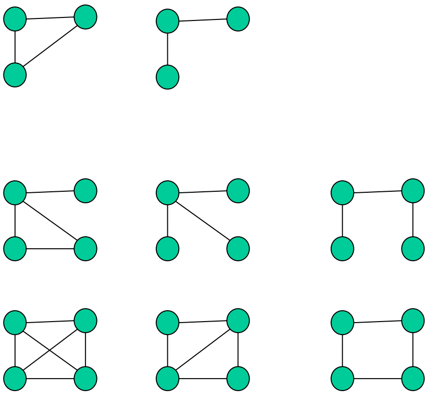
Superfamilies of Evolved and Designed Networks,
Ron Milo *et al.*, Science 2004.



Correlation Matrics



Undirected Graphs



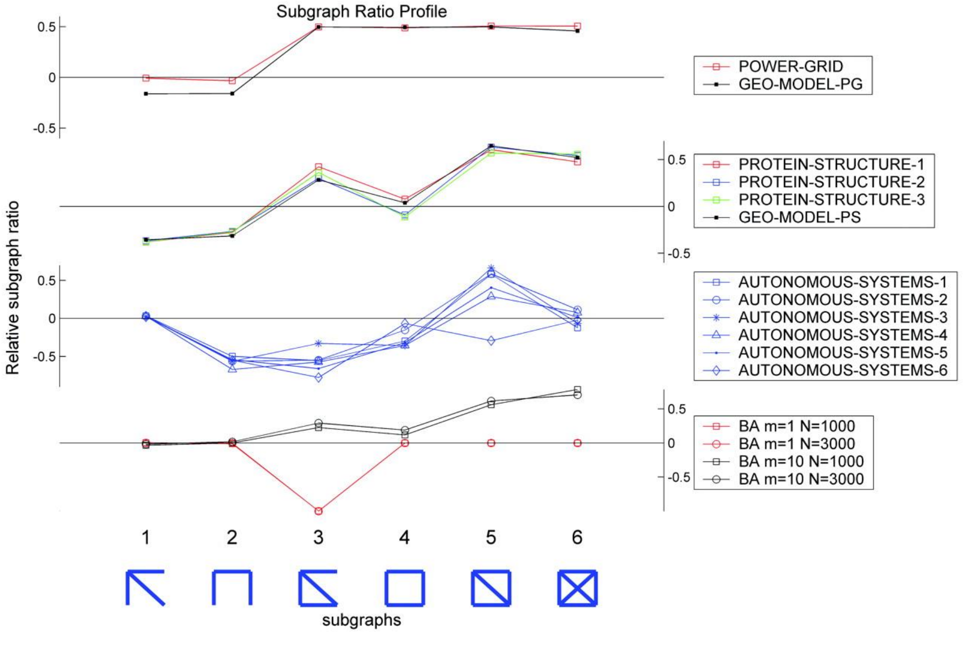
There are only 2 triads.
Use also tetrads.

Subgraph Ratio Profile (SRP)

- Unlike triads, the normalized Z scores of tetrads show a significant dependence on the network size.
- Therefore, instead of an SP based on Z scores, we use the abundance of each subgraph i relative to random networks:
  $\Delta_i = (N_{\text{real}_i} - \langle N_{\text{rand}_i} \rangle) / (N_{\text{real}_i} + \langle N_{\text{rand}_i} \rangle + \varepsilon)$

ε ensures that |Δ| is not misleadingly large when the subgraph appears very few times in both. Here, ε = 4.

$$SRP_i = \frac{\Delta_i}{\sqrt{\sum_{k=1}^{n} {\Delta_i}^2}}$$



Comparing Networks

- Given two networks how to compare there are identical? Similar?
  - Measure degree dist., CC, graphlet distribution
- Good to show that two graphs are different
- New definition
  - Degree distributions: how many nodes have k edges attached to them
  - Graphlet distribution: how many nodes have graphlet Y attached to them

[Pržulj, *Bioinformatics* 2007]

Graphlets and Automorphism Orbits