Master Thesis

Subject:<<Motifs of Network Models>>

Student : George Argyris

Advisor: Dimitris Kugiumtzis

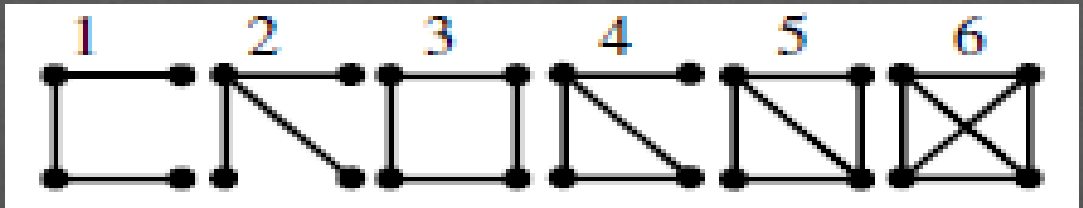
AUTH Thessaloniki December 2016

THE GOAL: Separate the concepts of randomness, smallworldness, scalefreeness

- Erdos-Renyi random networks (n, p)
- Watts-Strogatz small world networks (n, k, p_d)
- Barabasi-Albert scale free networks (n, G_i, m_0)
- Power law distribution scale free model (n, α)

Motifs

- Generally: statistically significant sub-graphs of a graph that occurs recurrently within the graph
- In our study, motifs are connected (tetrads) 4's nodes:



- Counting a graph's motifs:

-Work on the adjacency matrix

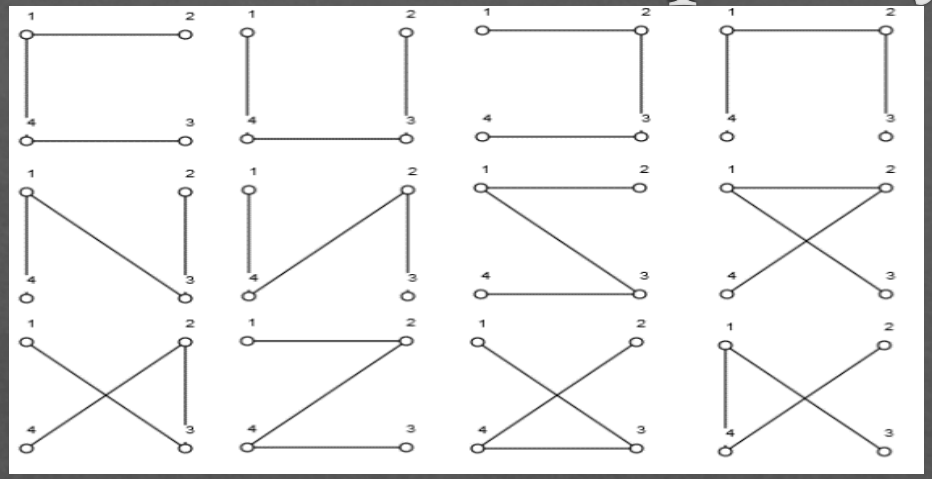
Examine all 4x4 sub-matrices of A $\left(\begin{matrix} n \\ 4 \end{matrix}\right) = n!/4!\cdot(n-4)! = n\cdot(n-1)\cdot(n-2)\cdot(n-3)/24$

-Symmetric graphs' problem: The algorithm was computationally complex because one motif is depicted by more than one adjacency matrices.

Solution: Same degree distribution for graphs of size: $n \leq 4 \Leftrightarrow$ isomorphic motifs

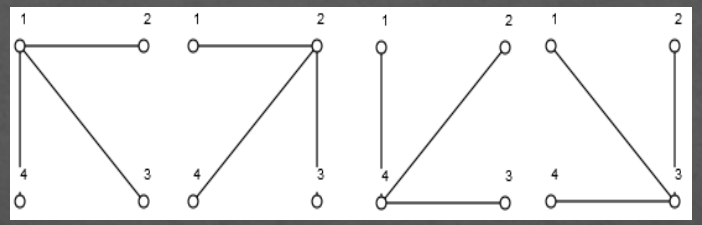
-I used Brain Connectivity Toolbox in order to further reduce computational complexity.

Motif 1 frequency in ER random graphs



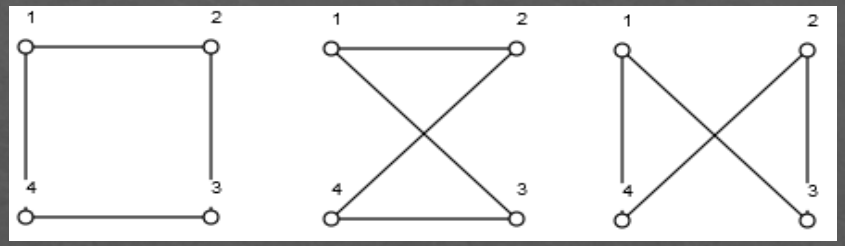
- 12 symmetries
- occurrence probability for each symmetry: $p^3 \cdot (1-p)^3$
- Motif's 1 frequency: $f_1 = 12 \cdot \binom{n}{4} \cdot p^3 \cdot (1-p)^3$

Motif 2 frequency in E-R random graphs



- 4 symmetries
- Probability of occurrence of each one of them: $p^3 \cdot (1-p)^3$
- Motif's 2 frequency: $f_2 = 4 \cdot (\blacksquare n@4) \cdot p^3 \cdot (1-p)^3$

Motif 3 frequency in E-R random graphs



- 3 symmetries
- Probability of occurrence of each one of them: $p^4 \cdot (1-p)^2$
- Motif's 3 frequency: $f\downarrow 2 = 3 \cdot (\blacksquare n@4) \cdot p\uparrow 4 \cdot (1-p)\uparrow 2$

Motif frequency of Erdos-Renyi random graphs

- Frequency of motif i:

$$f_i(n,p) = N_{mi} \cdot \binom{n}{4} \cdot p^{l} \cdot (1-p)^{6-l} \quad (1)$$

- Relative frequency of motif i:

$$F_i(n,p) = f_i(n,p) / \sum_{i=1}^{6} f_i(n,p)$$

- Motifs' relative frequency vector:

$$F = (F_1, F_2, F_3, F_4, F_5, F_6)$$

Highlight 1

Motifs' frequency comparison between ER and WS networks

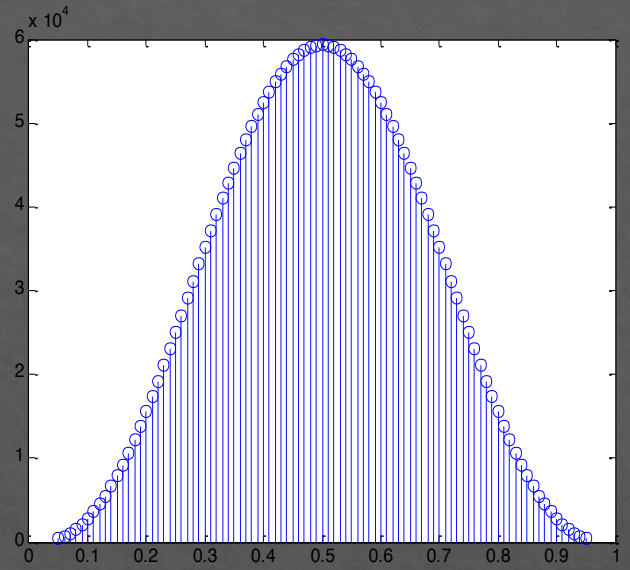
We counted the motifs' frequency for 25 networks of 25 nodes

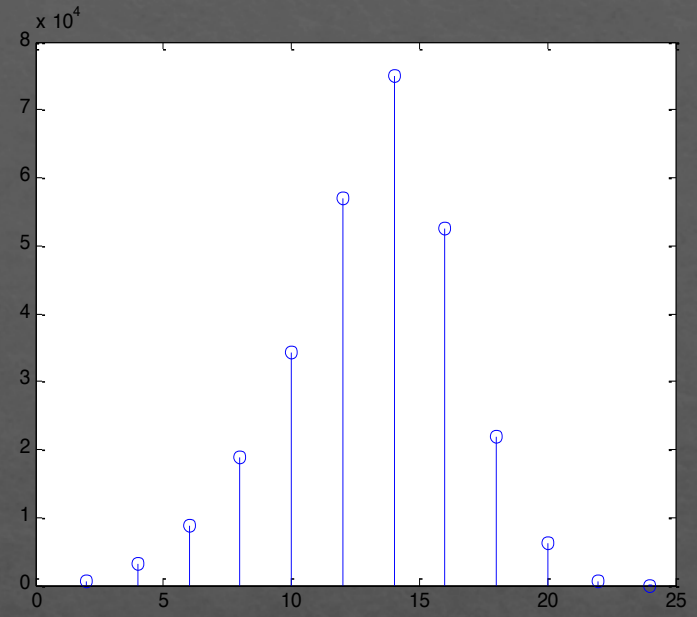
(random and small world ($p_d = 0$)) .

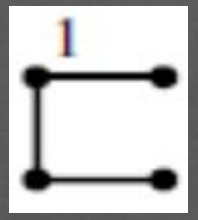
- In case of ER networks we constructed the graphs analytically, for different values of p, according to this function:

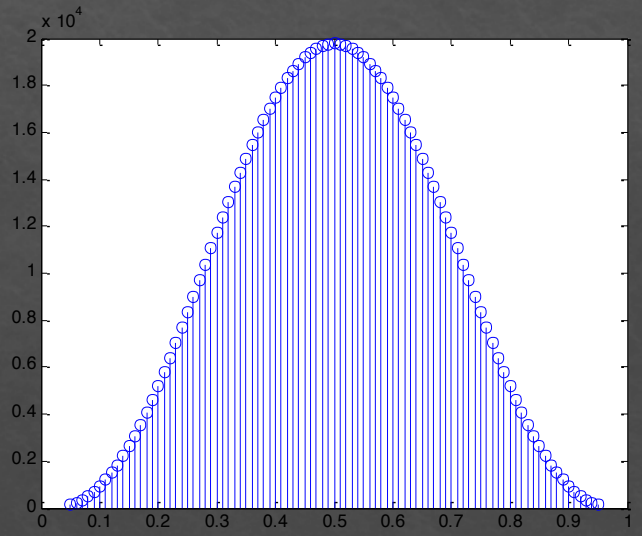
$$m\downarrow i\,(p)=25\cdot f\downarrow i\,(25,p)=25\cdot N\downarrow mi\cdot(\blacksquare n@4\ )\cdot p\uparrow l\cdot p\uparrow 6-l$$

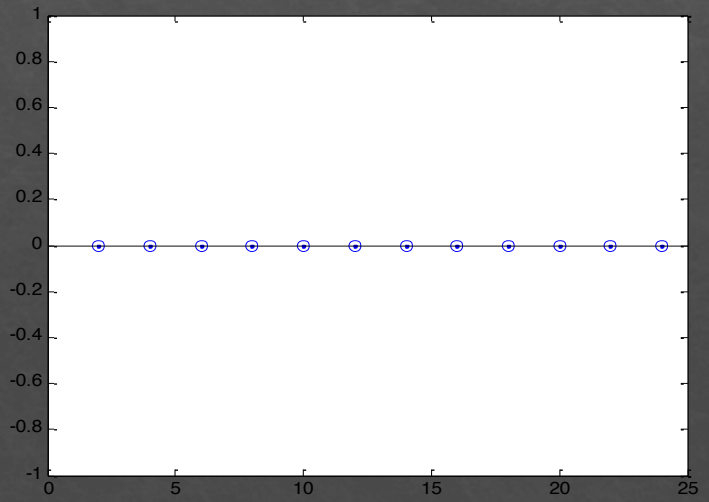
- In case of WS networks we constructed the graphs empirically, for different values of k and $p_d = 0$.

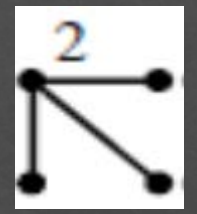

x 10^4
6
5
4
3
2
1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

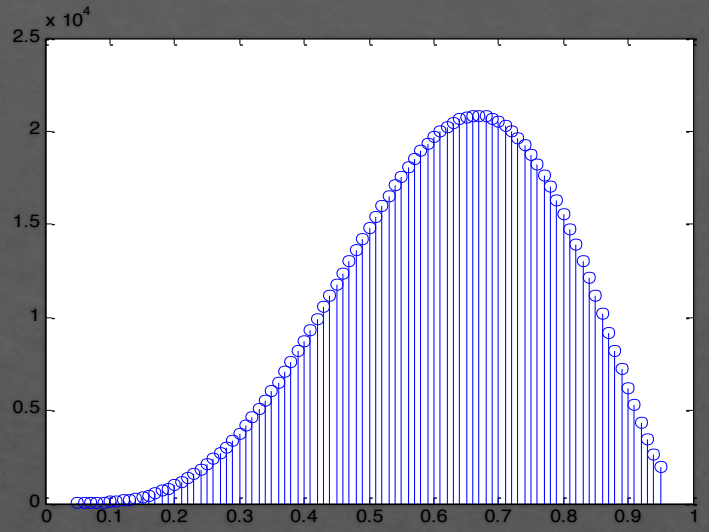

x 10^4
8
7
6
5
4
3
2
1
0
0 5 10 15 20 25

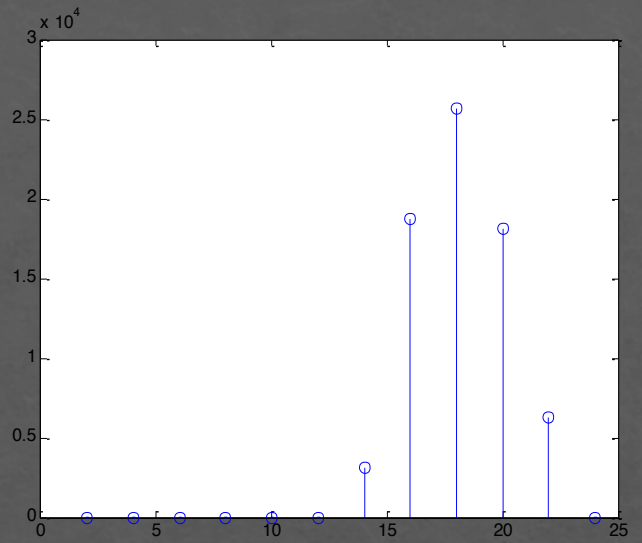

1

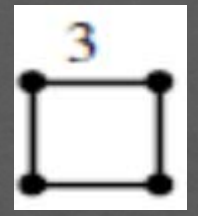

x 10^4
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

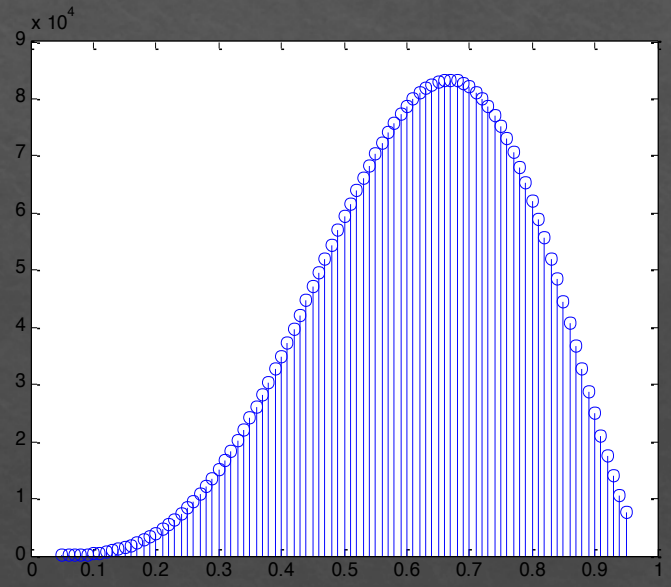

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0 5 10 15 20 25

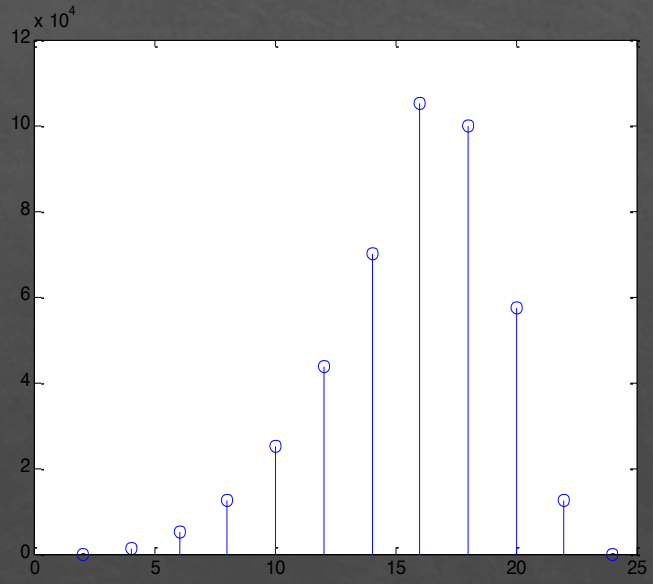

2

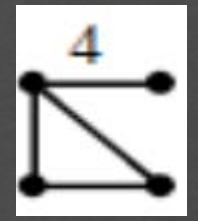

x 10^4
2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

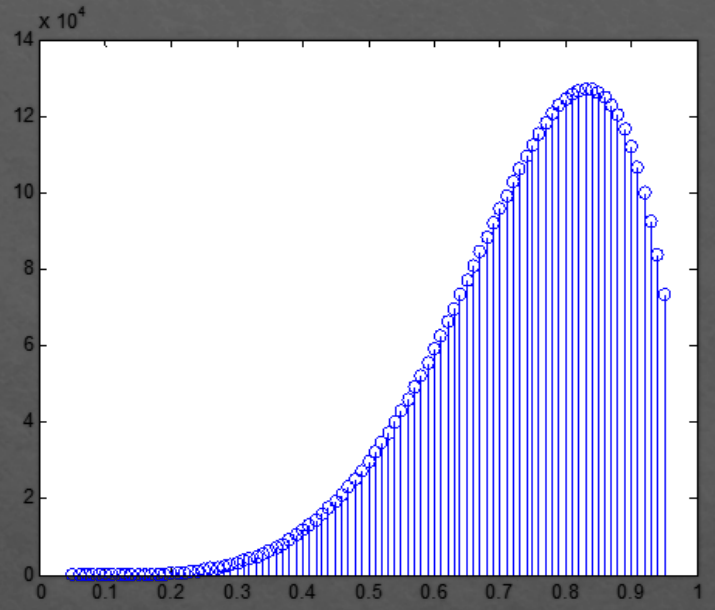

x 10^4
3
2.5
2
1.5
1
0.5
0
0 5 10 15 20 25

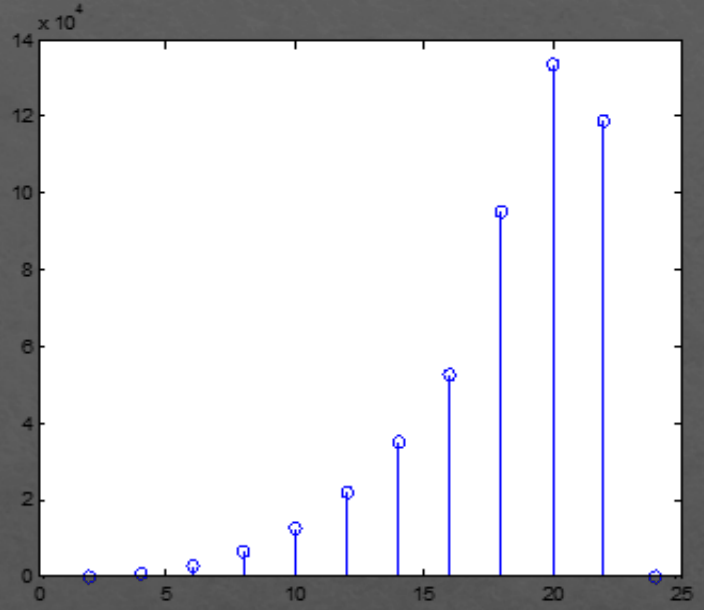

3

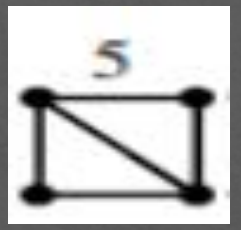

x 10^4
9
8
7
6
5
4
3
2
1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

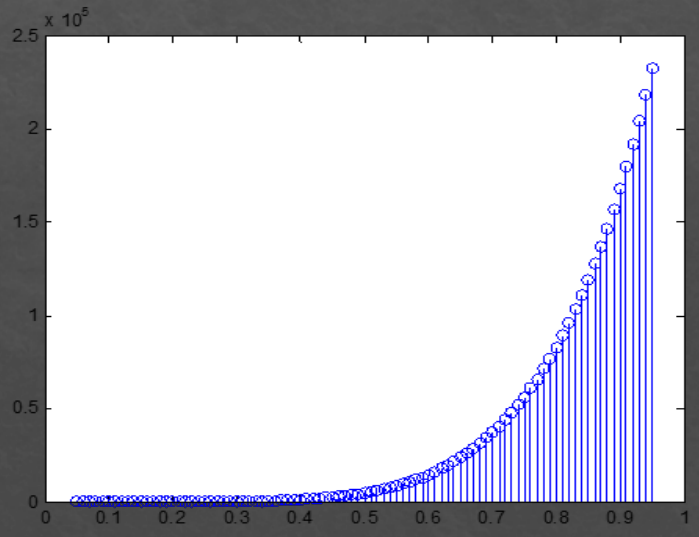

x 10^4
12
10
8
6
4
2
0
0 5 10 15 20 25

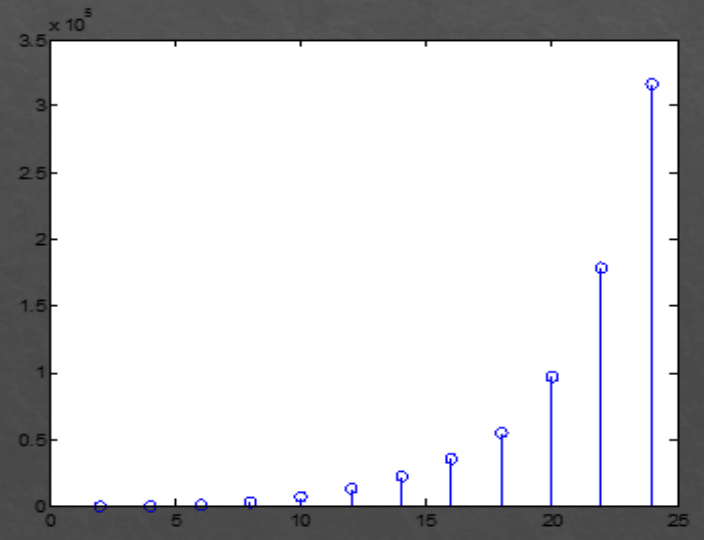

4


x 10^4
14
12
10
8
6
4
2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


x 10^4
14
12
10
8
6
4
2
0
0 5 10 15 20 25


5


x 10^5
2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


x 10^5
3.5
3
2.5
2
1.5
1
0.5
0
0 5 10 15 20 25


6

Highlights

- The smaller the euclidean distance between the motifs' relative frequency vector of an observed network and the expected motifs' relative frequency vector calculated previously the more random the network is.

- Maybe, we have same degree distributions for 5 of 6 motifs with different parameters independently the construction model.

- Motif 2 of great importance! As p of WS model approximates 1 the frequency of that motif tends to the expected value of a random one.

Problem-Method

- Suppose we have an observed network, could we find which model generates it? Or which model best approximates it?

Method

- Construction of the motifs' relative frequency vector for each one of the models and for each of their parameters as n=25, 50, 75.

  E-R: p=0.1, 0.2, …, 0.9

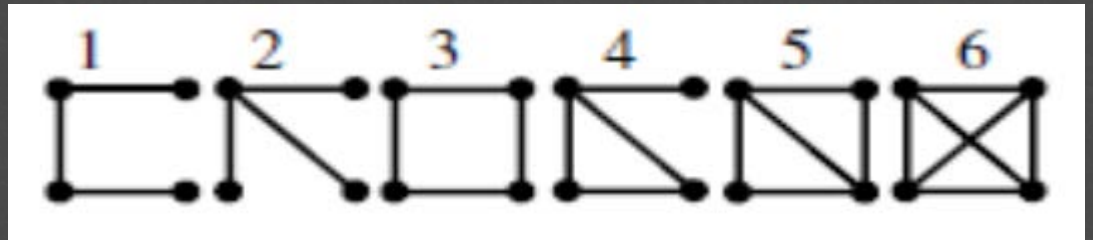
  W-S: k=2, 4, 6,…, p_d = 0.1, 0.2, …, 0.9

  B-A: m_i= each one of the initial motifs, i=1,…, 6, m_0=1, 2, 3, 4

  EXP: exp=2.1, 2.2, 2.3,…,2.9

- Finding the corresponding vector of the observed network.
- The model that generates the network is the one that minimizes the euclidean distance.

- We expect intuitively that:



Assessing the reliability of the method for 100 Erdos-Renyi graphs and for different fundamental scale free models

	RN 0.1	RN 0.2	RN 0.3	RN 0.4	RN 0.5	RN 0.6	RN 0.7	RN 0.8	RN 0.9
RN 0.1	3	1	0	0	0	0	0	0	0
RN 0.2	5	31	2	0	0	0	0	0	0
RN 0.3	0	7	56	4	0	0	0	0	0
RN 0.4	0	0	2	48	0	0	0	0	0
RN 0.5	0	0	0	0	50	1	0	0	0
RN 0.6	0	0	0	0	1	61	0	0	0
RN 0.7	0	0	0	0	0	1	50	2	0
RN 0.8	0	0	0	0	0	0	1	66	0
RN 0.9	0	0	0	0	0	0	0	1	51
WS 0.8,3	0	8	8	0	0	0	0	0	0
WS 0.8,4	0	0	6	1	0	0	0	0	0
WS 0.8,5	0	0	2	14	0	0	0	0	0
WS 0.8,6	0	0	0	4	12	0	0	0	0
WS 0.9,2	5	4	0	0	0	0	0	0	0
WS 0.9,3	0	6	2	0	0	0	0	0	0
WS 0.9,4	0	0	4	2	0	0	0	0	0
WS 0.9,5	0	0	1	16	0	0	0	0	0
WS 0.9,6	0	0	0	2	22	0	0	0	0
WS 0.9,7	0	0	0	0	4	19	0	0	0
WS 0.9,8	0	0	0	0	0	3	17	0	0
WS 0.9,9	0	0	0	0	0	0	7	11	0
WS 0.9,10	0	0	0	0	0	0	0	12	9
WS 0.9,11	0	0	0	0	0	0	0	0	38
EXP 2.1	0	4	0	0	0	0	0	0	0
EXP 2.2	9	15	0	0	0	0	0	0	0
EXP 2.3	6	0	0	0	0	0	0	0	0
EXP 2.4	8	1	0	0	0	0	0	0	0
EXP 2.5	15	0	0	0	0	0	0	0	0
EXP 2.6	0	0	0	0	0	0	0	0	0
EXP 2.7	0	0	0	0	0	0	0	0	0
EXP 2.8	0	0	0	0	0	0	0	0	0
EXP 2.9	5	0	0	0	0	0	0	0	0

	WS 0, 1	WS 0, 2	WS 0, 3	WS 0, 4	WS 0, 5	WS 0, 6	WS 0, 7	WS 0, 8	WS 0, 9	WS 0, 10	WS 0, 11
WS 0, 1	100	0	0	0	0	0	0	0	0	0	0
WS 0, 2	0	100	0	0	0	0	0	0	0	0	0
WS 0, 3	0	0	100	0	0	0	0	0	0	0	0
WS 0, 4	0	0	0	100	0	0	0	0	0	0	0
WS 0, 5	0	0	0	0	100	0	0	0	0	0	0
WS 0, 6	0	0	0	0	0	100	0	0	0	0	0
WS 0, 7	0	0	0	0	0	0	100	0	0	0	0
WS 0, 8	0	0	0	0	0	0	0	100	0	0	0
WS 0, 9	0	0	0	0	0	0	0	0	100	0	0
WS 0, 10	0	0	0	0	0	0	0	0	0	100	0
WS 0, 11	0	0	0	0	0	0	0	0	0	0	100

	WS 0.1, 1	WS 0.1, 2	WS 0.1, 3	WS 0.1, 4	WS 0.1, 5	WS 0.1, 6	WS 0.1, 7	WS 0.1, 8	WS 0.1, 9	WS 0.1, 10	WS 0.1, 11
WS 0.1, 1	76	0	0	0	0	0	0	0	0	0	0
WS 0.1, 2	0	69	0	0	0	0	0	0	0	0	0
WS 0.1, 3	0	0	55	0	0	0	0	0	0	0	0
WS 0.1, 4	0	0	9	63	0	0	0	0	0	0	0
WS 0.1, 5	0	0	0	7	71	0	0	0	0	0	0
WS 0.1, 6	0	0	0	0	0	81	0	0	0	0	0
WS 0.1, 7	0	0	0	0	0	0	79	0	0	0	0
WS 0.1, 8	0	0	0	0	0	0	0	69	0	0	0
WS 0.1, 9	0	0	0	0	0	0	0	0	57	0	0
WS 0.1, 10	0	0	0	0	0	0	0	0	0	52	0
WS 0.1, 11	0	0	0	0	0	0	0	0	0	0	29
WS 0, 1	10	0	0	0	0	0	0	0	0	0	0
WS 0, 2	0	16	0	0	0	0	0	0	0	0	0
WS 0, 3	0	0	15	0	0	0	0	0	0	0	0
WS 0, 4	0	0	0	14	0	0	0	0	0	0	0
WS 0, 5	0	0	0	0	9	0	0	0	0	0	0
WS 0, 6	0	0	0	0	0	10	0	0	0	0	0
WS 0, 7	0	0	0	0	0	0	8	0	0	0	0
WS 0, 8	0	0	0	0	0	0	0	11	0	0	0
WS 0, 9	0	0	0	0	0	0	0	0	17	0	0
WS 0, 10	0	0	0	0	0	0	0	0	0	24	0
WS 0, 11	0	0	0	0	0	0	0	0	0	0	40
WS 0.2, 1	11	0	0	0	0	0	0	0	0	0	0
WS 0.2, 2	0	13	0	0	0	0	0	0	0	0	0
WS 0.2, 3	0	0	20	0	0	0	0	0	0	0	0
WS 0.2, 4	0	0	0	15	0	0	0	0	0	0	0
WS 0.2, 5	0	0	0	0	20	0	0	0	0	0	0
WS 0.2, 6	0	0	0	0	0	9	0	0	0	0	0
WS 0.2, 7	0	0	0	0	0	0	13	0	0	0	0
WS 0.2, 8	0	0	0	0	0	0	0	19	0	0	0
WS 0.2, 9	0	0	0	0	0	0	0	0	26	0	0
WS 0.2, 10	0	0	0	0	0	0	0	0	0	23	0
WS 0.2, 11	0	0	0	0	0	0	0	0	0	0	24

Why method does not work well

- Method does not work well because:

1. Networks constructed by one model are classified to close models.

   Ex. ER(p) networks are mapped ER(p'), $p' = p \pm 0.1$

   WS(k, p) are mapped to WS(k', p') with $p' = p \pm 0.1$ and $k' = k \pm 2$

2. Density is the most critical parameter that affects the relative frequency of a vector.

Solution

If parameter density of the observed network is known, we could calculate some of the parameters of the models. So, the models that generate an observed graph become less, and the method more accurate.

The parameter calculation

- Concerning density d from an observed network, we could calculate some of the parameters of the models as below:

Erdos-Renyi: p = d

Watts-Strogatz:

$$d = GraphOrder/MaximumOrder = n \cdot k/2 \ / n \cdot (n-1)/2 \ = k/n-1 \Rightarrow$$
$$k = d \cdot (n-1)$$

Barabasi-Albert:

$$m\downarrow 0 = (d \cdot (n-1) \cdot n/2 \ -4)/n-4$$

Success rate of the method for known density, n=50

	RN.0.1	RN.0.2	RN.0.3	RN.0.4	RN.0.5	RN.0.6	RN.0.7	RN.0.8	RN.0.9
RN	74	75	83	85	96	97	95	99	95
WS.0.0	0	0	0	0	0	0	0	0	0
WS.0.1	0	0	0	0	0	0	0	0	0
WS.0.2	0	0	0	0	0	0	0	0	1
WS.0.3	0	0	0	0	0	0	0	0	0
WS.0.4	1	1	0	0	0	0	0	0	0
WS.0.5	2	0	0	0	0	0	0	0	0
WS.0.6	5	2	0	0	0	0	0	0	1
WS.0.7	3	4	1	0	0	0	0	0	1
WS.0.8	4	2	3	2	1	0	0	0	0
WS.0.9	11	16	13	13	3	3	5	1	2
BA.m1	0	0	0	0	0	0	0	0	0
BA.m2	0	0	0	0	0	0	0	0	0
BA.m3	0	0	0	0	0	0	0	0	0
BA.m4	0	0	0	0	0	0	0	0	0
BA.m5	0	0	0	0	0	0	0	0	0
BA.m6	0	0	0	0	0	0	0	0	0

Success rate of the method for standard density n=50

	WS.0.9,1	WS.0.9,2	WS.0.9,3	WS.0.9,4	WS.0.9,5	WS.0.9,6	WS.0.9,7	WS.0.9,8	WS.0.9,9	WS.0.9,10	WS.0.9,11	WS.0.9,12
RN	20	16	7	3	5	5	3	7	4	5	7	5
WS.0.0	0	0	0	0	0	0	0	0	0	0	0	0
WS.0.1	0	0	0	0	0	0	0	0	0	0	0	0
WS.0.2	0	0	0	0	0	0	0	0	0	0	0	0
WS.0.3	1	0	0	0	0	0	0	0	0	0	0	0
WS.0.4	5	1	0	0	0	0	0	0	0	0	0	0
WS.0.5	18	6	1	0	0	0	0	0	0	0	0	0
WS.0.6	27	10	12	5	7	1	2	0	0	0	0	0
WS.0.7	10	11	11	14	18	9	13	11	12	14	8	1
WS.0.8	2	26	9	15	19	25	26	23	21	21	26	20
WS.0.9	17	30	60	63	51	60	56	59	63	60	59	74
BA.m1	0	0	0	0	0	0	0	0	0	0	0	0
BA.m2	0	0	0	0	0	0	0	0	0	0	0	0
BA.m3	0	0	0	0	0	0	0	0	0	0	0	0
BA.m4	0	0	0	0	0	0	0	0	0	0	0	0
BA.m5	0	0	0	0	0	0	0	0	0	0	0	0
BA.m6	0	0	0	0	0	0	0	0	0	0	0	0

Success rate of the method for standard density n=50

	BA.m1.1	BA.m1.2	BA.m1.3	BA.m1.4	BA.m2.1	BA.m2.2	BA.m2.3	BA.m2.4
RN	0	4	2	0	1	1	0	0
WS.0.0	0	0	0	0	0	0	0	0
WS.0.1	0	0	0	0	0	0	0	0
WS.0.2	0	0	0	0	0	0	0	0
WS.0.3	0	0	0	0	0	0	0	0
WS.0.4	0	0	0	0	0	0	0	0
WS.0.5	0	0	0	0	0	0	0	0
WS.0.6	0	0	0	0	0	0	0	0
WS.0.7	0	0	0	0	0	0	0	0
WS.0.8	0	0	0	0	0	0	0	0
WS.0.9	0	0	0	0	0	0	0	0
BA.m1	58	8	49	42	46	5	48	46
BA.m2	17	3	0	13	17	0	0	6
BA.m3	0	37	0	5	0	49	0	2
BA.m4	0	0	9	0	0	0	4	1
BA.m5	14	48	32	11	30	45	34	11
BA.m6	11	0	8	29	6	0	14	34

	BA.m3.1	BA.m3.2	BA.m3.3	BA.m3.4	BA.m4.1	BA.m4.2	BA.m4.3	BA.m4.4
RN	0	6	3	0	0	1	0	2
WS.0.0	0	0	0	0	0	0	0	0
WS.0.1	0	0	0	0	0	0	0	0
WS.0.2	0	0	0	0	0	0	0	0
WS.0.3	0	0	0	0	0	0	0	0
WS.0.4	0	0	0	0	0	0	0	0
WS.0.5	0	0	0	0	0	0	0	0
WS.0.6	0	0	0	0	0	0	0	0
WS.0.7	0	0	0	0	0	0	0	0
WS.0.8	0	0	0	0	0	0	0	0
WS.0.9	0	0	0	0	0	0	0	0
BA.m1	41	6	49	49	36	11	49	35
BA.m2	28	0	0	13	11	3	0	11
BA.m3	0	38	0	4	0	34	0	1
BA.m4	0	0	3	3	0	0	8	5
BA.m5	17	50	39	4	28	51	27	11
BA.m6	14	0	6	27	25	0	16	35

Application to real data

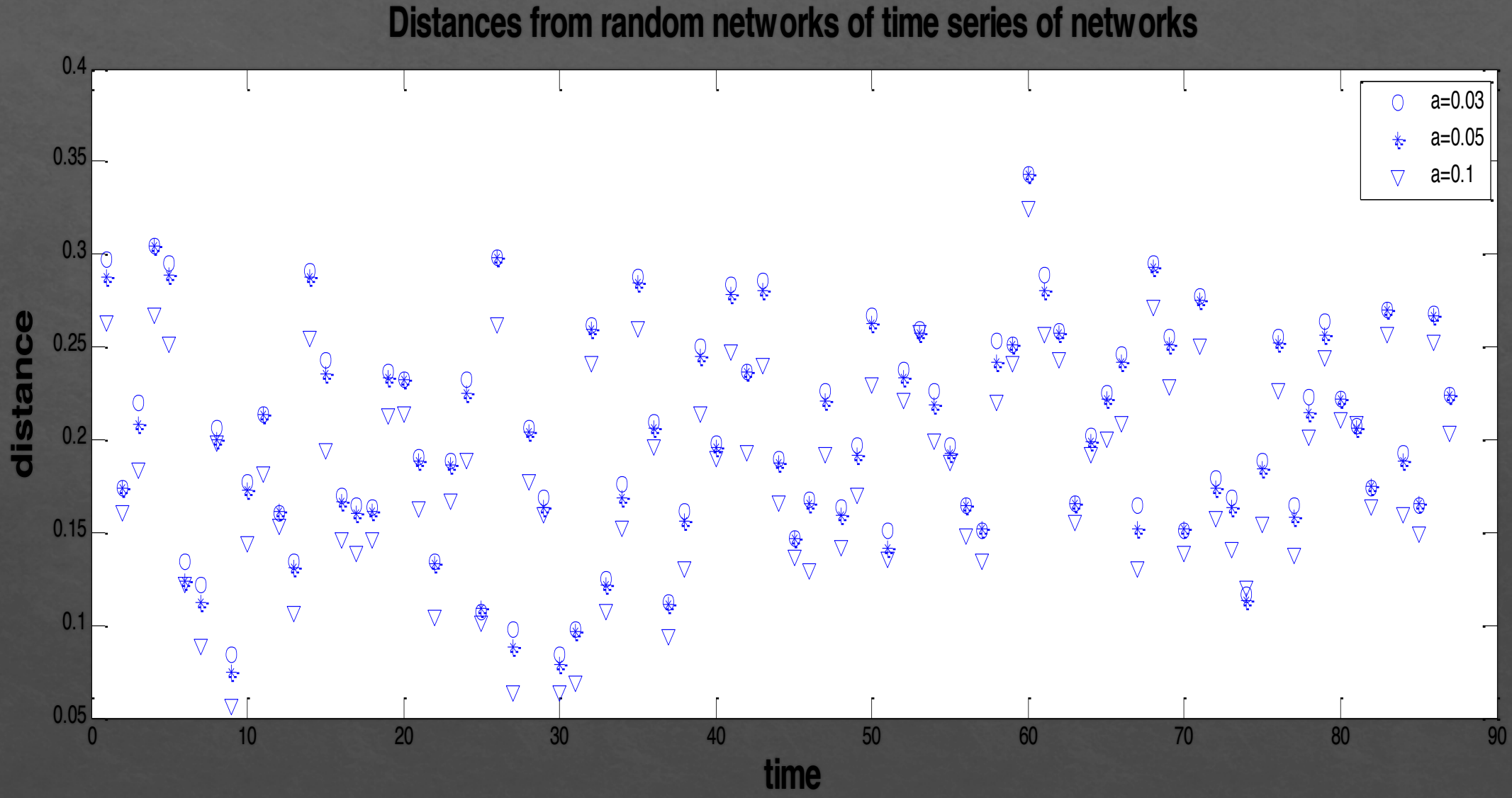
- The Morgan Sanley Capital International's (MSCI) dataset contains the capitalization indices for 55 markets.
- It contains 1305 daily returns for each market from 5 March 2004 to 5 March 2009.
- We separate the dataset into 87 and 29 subdatasets each one of them contains 15 and 45 returns respectively in chronological order. For each one of the subdatasets we have an association network which is constructed by the correlation of the time series.

Statistical procedures used

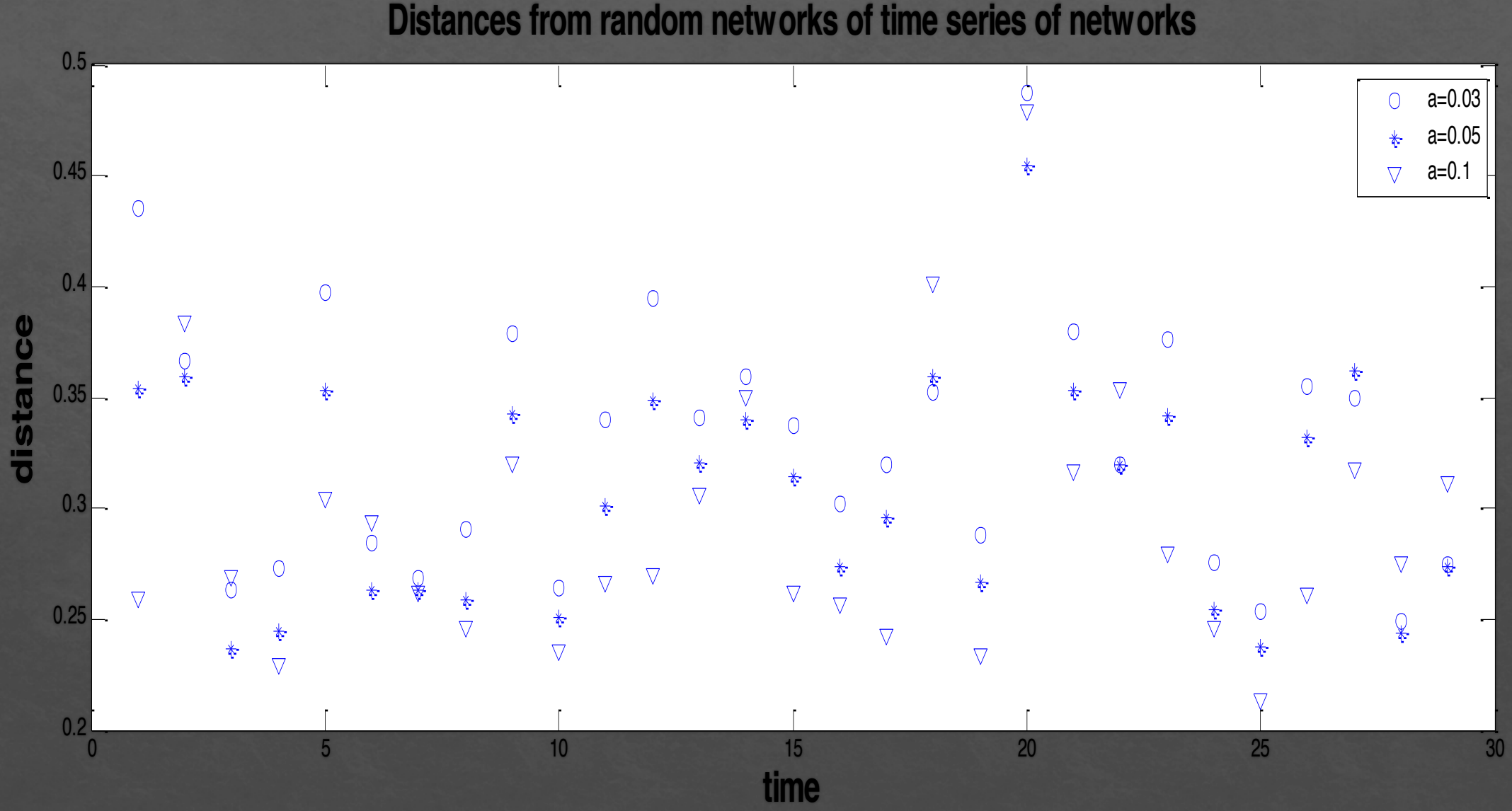
For the statistical analysis of the multivariate time series we used:

- For each one of the subdataset and for each one of the time series a prewhiten method.
- A significance test for finding the statistically significant cross correlation. In that test there is a significance level a. As a gets large, the cross-correlation that are statistically significant are getting large. Also, a determines the density d of the association network.

The randomness indicator for the association networks for the case of 87 networks



The randomness indicator for the association networks for the subdataset of 29 networks



Conclusion

Statement 1.

Let G be an ER(p) graph and G_1, G_2 be subgraphs of same order and size.

The frequency of G_1 is higher than the frequency of G_2 if and only if G_1 has more symmetries than G_2.

Statement 2.

Subgraph G achieves highest frequency in an ER(p) graph when $p = d_G$.

Future research activities

- Find the motifs of a WS(n, k, 0) model analytically.
- Find which motif triad better recognizes the model and make a 3D representation.
  The statement 2 may help to this direction.
- Application to other multivariate time series.
- Instead of euclidean distance, we could train a perceptron which works if-f the data are linearly separable and we compare only 2 models.
- Same analysis for motifs of size 5 is feasible.
- Extendable to directed networks.
- The method might be extendable to other network models.