Statistical Mechanics of Complex Systems

Application of maximum entropy and network analysis

Anna Braghetto - 1205200; Davide Maniscalco - 1212063

Presentation of the dataset

The initial dataset contained information about 368122 trees in 50-hectare plot at Barro Colorado Island, Panama. After discarding died trees (159682) and not avilable data-lines (53), we obtained a dataset containing 208387 trees of 299 different species.

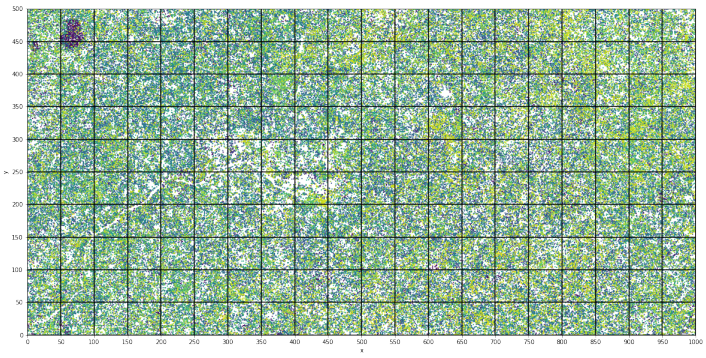

Figure 1: The Barro Colorado Forest with its subdivision in subplots

To perform the analysis the area is divided into 200 subplots of 0.25 hectare each, and, for each one, the abundances (i.e. the number of alive trees) are calculated for each species.

The abundance matrix (subplots and species) is shown in Fig. 2.

For most the species and subplot the abundance is goes from 0 to 100: the differences are not visible; just to make them clearer, we plot the matrix again with a cut on the abundance to 20 in Fig. 3.

The "presence" of a species in an arbitrary subplot is defined as follows:

$$p_i^\alpha = \begin{cases} +1 & \text{If species } i \text{ is present in the subplot } \alpha \\ 0 & \text{If species } i \text{ is not present in the subplot } \alpha \end{cases}$$

For each species $i = (1, ..., s)$, where s is the number of species (i.e. $s = 299$), we calculate the mean presence p_i, averaging on the 200 subplots:

$$p_i = \frac{1}{200} \sum_{\alpha=1}^{200} p_i^\alpha$$

In Fig.4 we show the barplot of the average presences of the different species.

Maximum entropy model 1

A first maximum entropy model is defined as follows:

$$P(\vec{\sigma}, \vec{\lambda}) = \frac{1}{Z_{\text{tot}}} e^{\sum_{i=1}^{s} \lambda_i \sigma_i}$$

where λ_i are the Langrangian parameters that satisfy the following constrain:

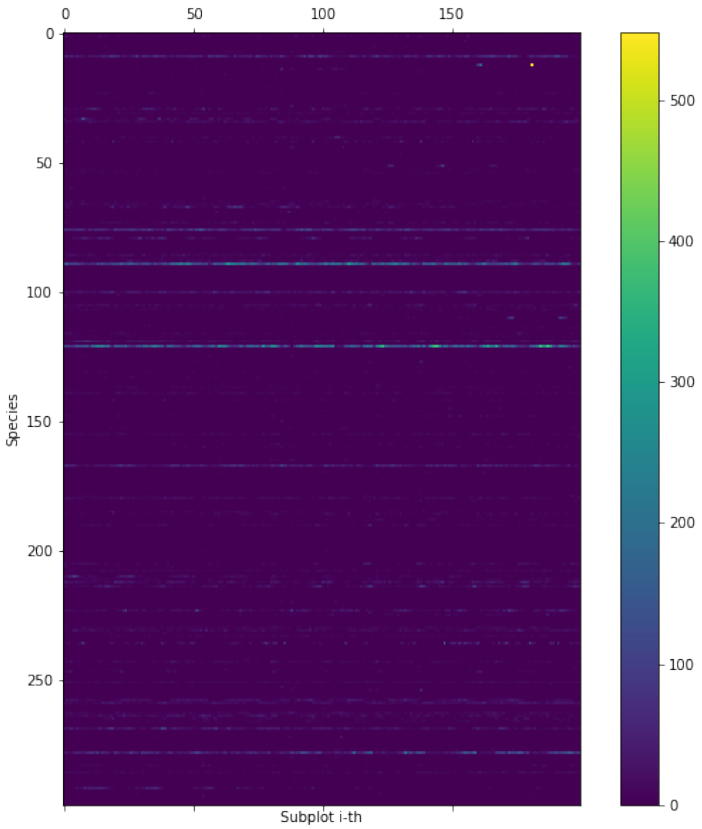

Figure 2: Abundances for each species in each subplot.

$$\langle \sigma_i \rangle_{emp} = m_i$$

with

$$m_i = 2p_i - 1$$

where p_i is the average presence i-th computed above. The single presence-probabilities are independent, in fact the "presence" probability for a species i is

$$P_i(\lambda_i, \sigma_i) = \frac{1}{Z} e^{\lambda_i \sigma_i}$$

where λ_i is the i-th Lagrangian parameter, and Z is the partition function. Since $\sigma_i = \pm 1$, the partition function Z can be calculated expanding the sum:

$$Z = \sum_{\sigma_i = \pm 1} e^{\lambda_i \sigma_i} = 2\cosh(\lambda_i)$$

and the average $\langle \sigma_i \rangle_{model}$ is then computed as follows:

$$\langle \sigma_i \rangle_{\text{model}} = \frac{1}{Z} \frac{\partial Z}{\partial \lambda_i} = \tanh(\lambda_i) \, .$$

By imposing the constraint $\langle \sigma_i \rangle_{\text{model}} = \langle \sigma_i \rangle_{emp}$, it is possible to find the Langrangian parameters as

$$\lambda_i = arctanh(\langle \sigma_i \rangle_{emp}) \, ;$$

in Fig.5 is shown the histogram of the λ coefficients.

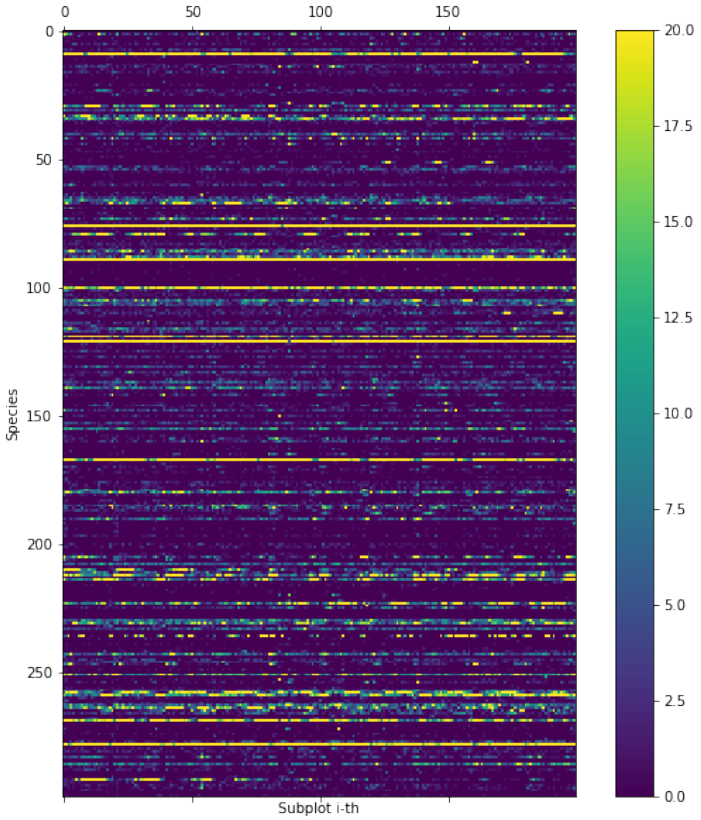

Figure 3: Abundances for each species in each subplot with a max abundance set to 20.

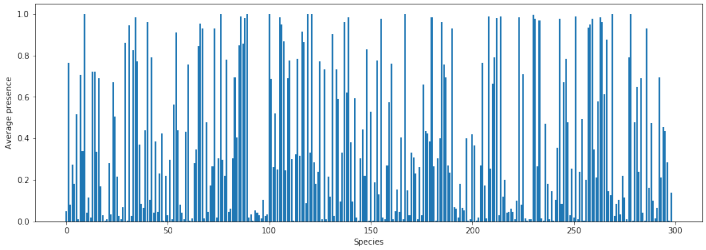

Figure 4: Average presence of each species

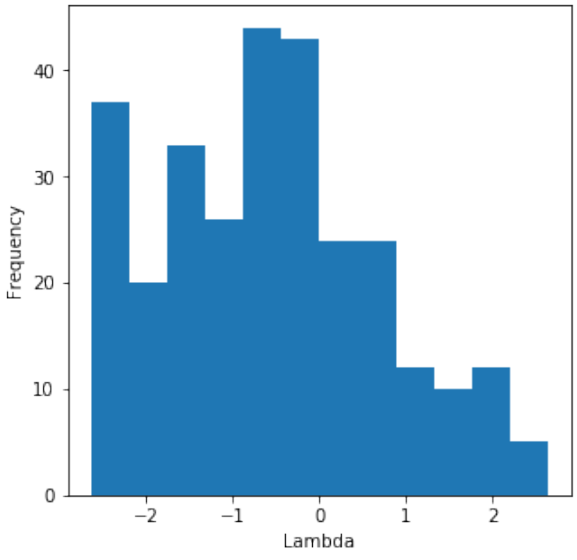

Figure 5: Histogram of λ coefficients for MaxEnt1 model; infinite values are not present.

Some of the Langrange parameters are infinite: they are relative to species present in *all* the subplots, i.e. these species have probability $P = 1$ to be present. λs are spread around zero: some species are even more present than others and by leaving out the infinite values, we find for the mean and the standard deviation:

$$\langle\vec{\lambda}\rangle \simeq -0.56 \qquad \text{StdDev}(\vec{\lambda}) \simeq 1.277$$

Maximum entropy model 2

A more complicated maximum entropy model is considered: it is described by the Hamiltonian:

$$H = -\sum_{j=1}^{s} \lambda_j \sigma_j - \frac{k}{s} \sum_{j=1}^{s} \sigma_j$$

and by the constraints:

$$\langle\sigma_i\rangle_{model} = \langle\sigma_i\rangle_{emp}$$

$$\left\langle \left(\sum_{j=1}^{s} \sigma_j\right)^2 \right\rangle_{model} = \left\langle \left(\sum_{j=1}^{s} \sigma_j\right)^2 \right\rangle_{emp} .$$

Unfortunately, it is not possible to obtain an analytical form for the 300 Lagrangian parameters (k and λ_i): in order to compute them we perform a simulation.

The Metropolis algorithm is used to simulate the configurations ($\sigma_{\{1,\dots,S\}} = \pm 1$) in order to obtain the equilibrium configuration by minimizing the energy of the system.

At first, a random initialization of the parameters is performed. Assuming that the lambdas follow a normal distribution, we extract them randomly from a Gaussian with mean and standard deviation set to the values computed at the previous point for the first maximum entropy model, while k is initialized at 1.5. With these parameters, the initial energy of the system is computed.

At each step of Metropolis, a random spin is flipped, leading to a new state with energy E': this new state will be accepted with probability $P = min(1, e^{-(E'-E)})$. Then the process is iterated a desired number of times.

To check the convergence (i.e. the achievement of an equilibrium configuration), two first simulations are run, with 10000 and 100000 iterations. The output are shown in Fig.6.

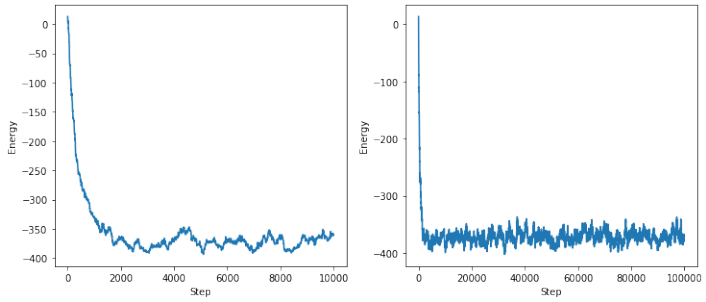

Figure 6: Energy evolution for the first two Metropolis simulations (10000 and 100000 iterations).

The algorithm converges to the minimum energy in 2000 steps.

Checked the convergence of Metropolis, the estimation of the Lagrangian parameters is made with the *Gradient descend algorithm*.

The parameters are initialized as before and, at each step of the Gradient Descend (GD), the Metropolis algorithm with 2500 iterations is run where just the last 500 spin configurations

(i.e. the equilibrium configurations) are saved. From these configurations we calculate the mean values of our interest of the model defined above by the constrains. Then, the parameters are updated on the basis of the difference between the empirical means and the model means just calculated, i.e.:

$$\lambda_i \longleftarrow \lambda_i + \eta(\langle\sigma_i\rangle_{emp} - \langle\sigma_i\rangle_{model})$$

$$k \longleftarrow k + \eta\left(\left\langle\left(\sum_{j=1}^{s}\sigma_j\right)^2\right\rangle_{emp} - \left\langle\left(\sum_{j=1}^{s}\sigma_j\right)\right\rangle_{model}\right)^2 ,$$

The learning rate η is set to 0.5 and the number of GD iterations is set to 5000. In Fig. 7 is shown the evolution of the k parameter.

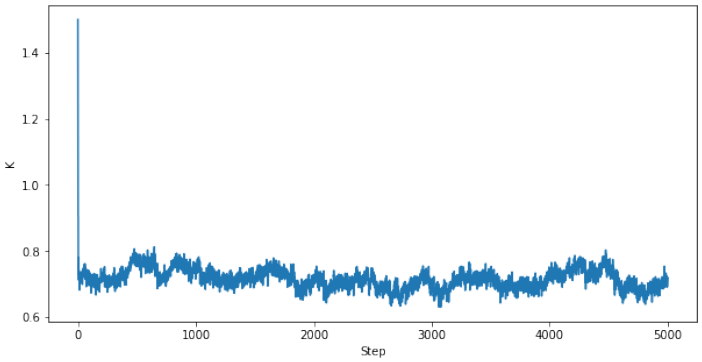

Figure 7: Evolution of the k parameter ($\eta = 0.5$,5000 iterations).

The k value immediately falls down 0.8 and fluctuates between 0.6 and 0.8 during all the simulation. We made the estimation of the k parameter averaging on the last 2500 iterations (half of the total) and associating as the error the standard deviation; so it is found:

$$k \simeq (0.706 \pm 0.028)\,.$$

For the estimation of the lambdas is made the same average procedure: the histogram obtained is shown in Fig.8.

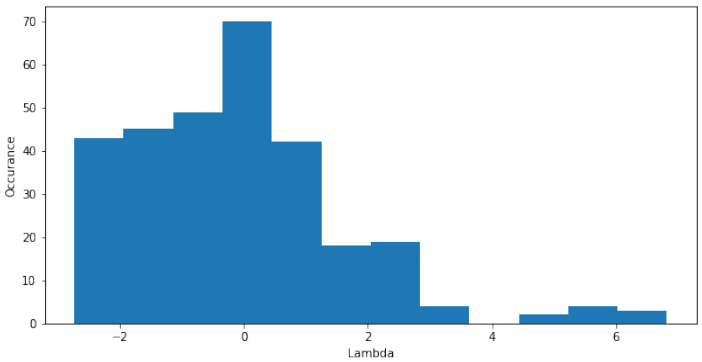

Figure 8: Histogram of the estimated values of the λ parameters. The estimation is made averaging the last 2500 values for each parameter.

Maximum entropy model 3

A last maximum entropy model is built as follows

$$P(\vec{x}, \vec{\mu}, M) = \frac{1}{Z} e^{(-\sum_i \mu_i x_i - \frac{1}{2}\sum_{ij} M_{ij} x_i x_j)}$$

where the variable x_i is the abundance of the species i, the Lagrangian parameters are the elements of the vector μ and of

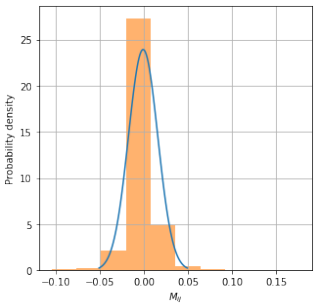

Figure 9: Histogram of the M_{ij} entries

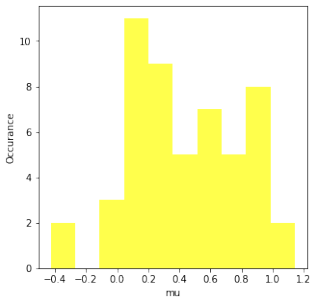

Figure 10: Histogram of the μ entries

the interaction matrix M (with $M_{ij} = M_{ji}$ and $M_{ii} = 0$: no self-interaction terms.)

The constraints are the following:

$$\langle x_i\rangle_{emp} = \langle x_i\rangle_{model}$$
$$\langle x_i x_j\rangle_{emp} = \langle x_i x_j\rangle_{model}\ ,$$

and, thank to the Gaussian approximation, the parameters μ_i are estimated as:

$$\mu_i = \sum_j -M_{ij}\langle x_j\rangle_{emp}$$

$$M_{ij}^{-1} = Cov(x_i, x_j) = \langle x_i, x_j\rangle_{emp} - \langle x_i\,.\rangle_{emp}\langle x_j\rangle_{emp}\,.$$

To perform the computation, we take into account only the most abundant species, discarding the ones that do not satisfy the inequality

$$\langle x_i\rangle_{emp} > \sigma_{x,i}\ :$$

we remain with 52 species, and we can obtain the Lagrangian parameters using the constraints just seen. In Fig.9 we show the histogram for the M_{ij} entries (where we set to zero the whole diagonal, that was containing self-interaction terms), while in Fig.10 the one for μ parameters.

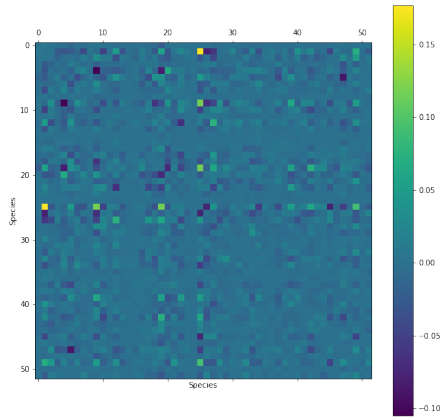

Figure 11: Interaction matrix M

The interaction matrix M is shown in Fig.11, and for its entries the following mean and variance are found:

$$\langle M_{ij}\rangle \simeq 8.83\cdot 10^{-4} \qquad \mathrm{var}(M_{ij}) \simeq 2.78\cdot 10^{-4}\,.$$

Thank to these, it is possible to compare the histogram with a normal distribution built with these mean and variance. It seems resonable that the M_{ij} entries follow a normal distribution almost symmetric and very narrow: few species interact strongly, many weekly, and the interaction can be either positive (we can maybe say "cooperative") or negative (non-cooperative).

Network analysis

As we can see from the histogram, the entries of M_{ij} follow a distribution very peaked on zero, and this leads us to conclude that there are a lot of small interactions. We want to leave out many, and for this purpose we set $M_{ij} = 0$ if $|M_{ij}| < \theta$ where θ is a parameter that can be chosen.

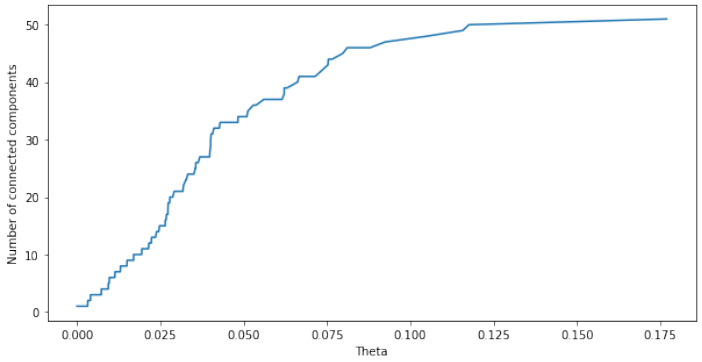

Figure 12: Number of connected components as a function of θ. It is possible to see the phase transition.

We make a graph W using the entries of M_{ij}.

In Fig.12 is shown the number of connected components as a function of θ. As we can see, we quickly move from a graph with a unique connected component to the null graph (52 isolated nodes). There's a kind of phase transition, and we found for the critical value:

$$\theta^* \simeq 0.003226 \ :$$

this is the greatest θ for which we have only one connected component. Let's call the corresponding graph W^*, shown in Fig.14, and study it. For diameter, global clustering coefficient, and assortativity degree, we have

$$D = 3 \qquad C \simeq 0.727 \qquad r \simeq -0.251$$

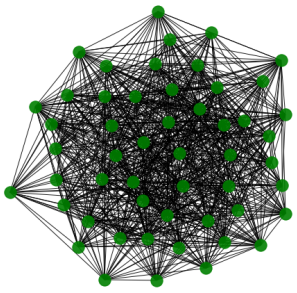

Figure 13: Erdos-Renyi graph

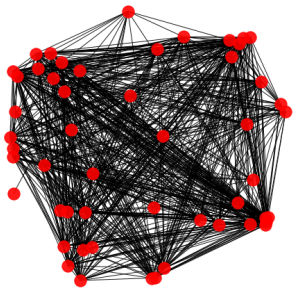

Figure 14: W^* graph

The diameter value is low, and furthermore the graph is also well clustered. We notice a slightly unassortative behaviour. There are a lot peripherical edges, and a few with high betweenness: probably there are few species that "bind" strongly with all the others.

At last, we want to study the comparison between the graph W^* and its corresponding Erdos-Renyi one. For this purpose, we generate a random ER graph with $N = 52$ nodes and with probability $p = \frac{\langle k \rangle}{N-1} \simeq 0.587$ (Fig. 13).

For diameter, global clustering coefficient, and assortativity degree, we have

$$D_{ER} = 2 \qquad C_{ER} \simeq 0.598 \qquad r_{ER} \simeq -0.018 \, .$$

These values are similar to the previous ones. Now, in Figg.15, 16, 17, 18, 19, 20 are made comparison between the two graphs on degreee, betweenness, and clustering.

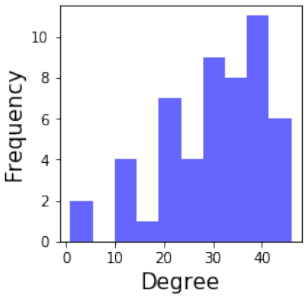

Figure 15: Histogram of degree distribution, W^*.

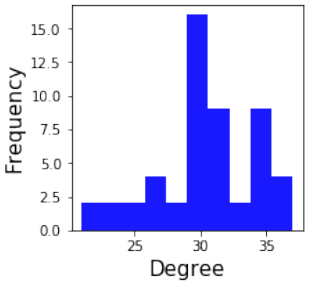

Figure 16: Histogram of degree distribution, Erdos-Renyi.

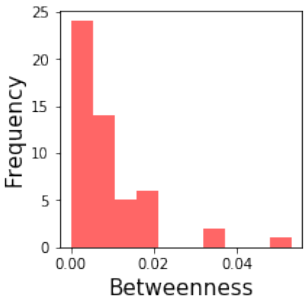

Figure 17: Histogram of betweenness centrality, W^*.

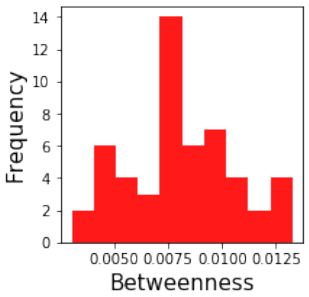

Figure 18: Histogram of betweenness centrality, Erdos-Renyi.

The average degrees, $\langle k \rangle \simeq 30$ and $\langle k_{ER} \rangle \simeq 30.35$ are very similar, while the degree distribution for the ER graph is more peaked around the mean.

The betweenness' histograms are, instead, very different: while the ER one gives an idea of symmetry, the W^* one is peaked at the beginning: less nodes have a higher 'importance' than the others.

A similar, but inverse, situation is found for the clusterings' histograms: while again the ER one gives an idea of symmetry, the W^* one is peaked at the end: the W^* graph is more clustered.

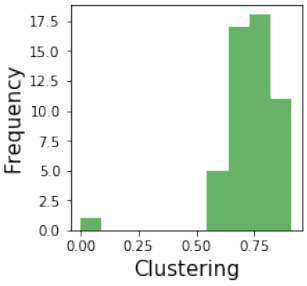

Figure 19: Histogram of clustering coefficients, W^*.

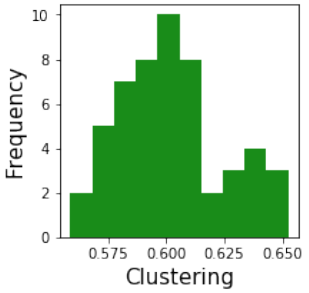

Figure 20: Histogram of clustering coefficients, Erdos-Renyi.