Analysis of the Barro Colorado Island dataset

Andrea Perin, 1206078 and Carlo Polato, 1205688

(Dated: April 7, 2019)

A presentation about the application of Max Ent techniques in the analysis of the Barro Colorado forest dataset.

I. TASKS 1/2: DATA EXTRACTION

Initially, the dataset contained information for both alive and dead trees. After filtering the dead ones out, we obtained that, over the whole plot, as of 2005, there were $S = 299$ different species of trees. The plot was then partitioned into $N = 200$ square subplots of side $l = 50$ m. Each of these was understood to be independent from the others; as such, a statistical analysis was carried over. In particular, for each subplot $j = 1, ..., N$, we calculated both the vector of *presences*, $\vec{p}_j \in \{0, 1\}^S$, setting the i-th entry to 1 if the i-th species was present in the subplot, and to 0 otherwise, and the vector of *abundances*, $\vec{x}_j \in \mathbb{N}^S$, the i-th entry of which corresponds to the number of trees of species i.

II. TASK 3: MAX ENT 1

The next step was to perform an analysis of the dataset by means of the *principle of maximum entropy (Max Ent)*. In the maximum entropy approach, the real (but unknown) distribution $P(G)$ is approximated by a model distribution $P(G)$ that maximizes the entropy. Using the graph formalism (that is, G denotes one possible graph belonging to the ensemble of graphs $\mathcal{G}$), given a probability distribution $P(G)$, we can write the entropy and the normalization condition as

$$S = -\sum_{G\in\mathcal{G}} P(G)\log P(G), \qquad \sum_{G\in\mathcal{G}} P(G) = 1.$$

Given a graph observable, $x_i(G)$, we can set a constraint on it by writing

$$\sum_{G\in\mathcal{G}} P(G)x_i(G) = \langle x_i(G)\rangle_{emp}, \quad i = 1, ..., S$$

where $\langle x_i(G)\rangle_{emp}$ is its empirical value, averaged over the available observations. Maximizing the entropy, together with the constraints (normalization and the additional S ones) with respect to the probability $P(G)$, gives an expression for $P(G)$ itself:

$$P(G) = \frac{1}{Z}\exp\left(\sum_{i=1}^{S}\lambda_i x_i(G)\right)$$

where λ_i are the Lagrange multipliers associated with the constraints on the observables and Z is the partition function, defined as

$$Z = \sum_{G\in\mathcal{G}}\exp\left(\sum_{i=1}^{S}\lambda_i x_i(G)\right)$$

From Z we can derive analytical expressions for the observables, as

$$\langle x_i\rangle_{emp} = \frac{\partial}{\partial\lambda_i}\log Z, \quad i = 1, ..., S$$

and then, by setting them to be equal to their empirical values, we can obtain the values of the parameters λ. Defining the graph hamiltonian as $H(G) = -\sum_{i=1}^{S}\lambda_i x_i(G)$, the partition function can be written as

$$Z = \sum_{G\in\mathcal{G}}\exp\left(-H(G)\right).$$

In the first model, the observables are the "magnetizations"

$$x_i(G) = 2p_i(G) - 1 = \sigma_i(G), \qquad i = 1, ..., S,$$

so that the constraints are

$$\sum_{G\in\mathcal{G}} P(G)\sigma_i(G) = \langle\sigma_i\rangle_{emp}, \qquad i = 1, ..., S$$

where G is to be understood as a possible configuration of magnetizations, and $\sigma_i(G)$ is such that, given a configuration (in our case, a subplot), its value is 1 if the species i is present and -1 otherwise. This results in the following expression for Z:

$$Z = \sum_{\{\sigma\}}\exp\left(\sum_{i=1}^{S}\lambda_i\sigma_i\right) \to Z = 2^S\prod_{i=1}^{S}\cosh(\lambda_i)$$

Then, for $i = 1, ...S$,

$$\langle\sigma_i\rangle_{model} = \frac{\partial}{\partial\lambda_i}\log Z \quad \to \quad \langle\sigma_i\rangle_{model} = \tanh(\lambda_i)$$

and imposing the constraints, we get that, for $i = 1, ..., S$,

$$\langle\sigma_i\rangle_{emp} = \langle\sigma_i\rangle_{model} \to \lambda_i = \tanh^{-1}(\langle\sigma_i\rangle_{emp})$$

The actual values of $\langle\sigma_i\rangle_{emp}$ are calculated as the element-wise average over the $N = 200$ subplots, given that, for each subplot, each σ is linked to the presence p_i by the relation $\sigma_i = 2p_i - 1$.
This approach, which is consistent in principle, leads to an (apparent) inconvenience: for species which are always present, their respective λs cannot be calculated, as $\tanh^{-1}(1)$ is undefined. However, the hamiltonian interpretation can aid us.
Let us consider a species which is always present in our

empirical dataset, so that its $\langle \sigma_i \rangle_{emp} = 1$. The Max Ent procedure will then yield a corresponding Lagrange multiplier $\lambda \to +\infty$. Then the probability of a configuration where the species i is not present is 0, as $e^{\lambda_i \sigma_i} = e^{-\infty} \to 0$. One may then argue that this particular species is "non-informative": the species *must* be present if we want a meaningful probability.
In terms of Shannon entropy, the event "the species i is present" is certain: its probability is 1 and, conversely, its entropy is 0. This is also consistent with the physical interpretation of the $\langle \sigma_i \rangle_{emp}$ as an Ising magnetization. An averaged magnetization of exactly 1 would only be possible in a 0 temperature regime, which in turn corresponds to no entropy.
In fig. 1, the values of the parameters are depicted. An arbitrary value of 10 is used in order to visualize the undefined values. In fig. 2, a histogram depicts the distribution of the (finite) $\{\lambda\}$. The compatibility of $\{\lambda\}$ with 0 is ≈ 0.44.

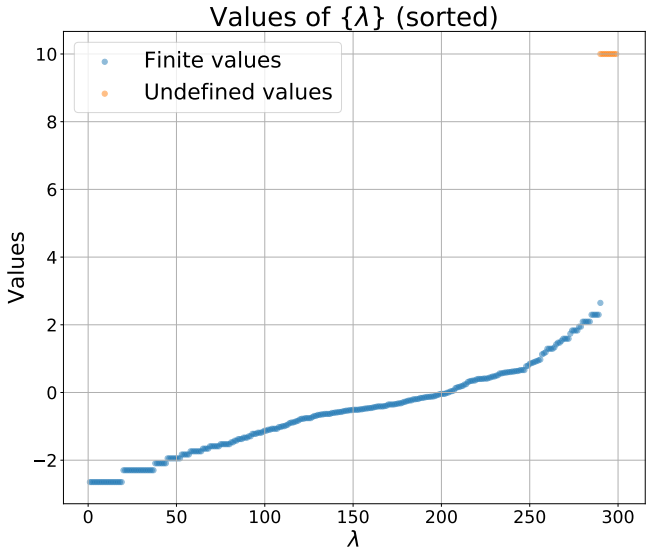

Figure 1: *The sorted values of the Lagrange parameters λ for the Max Ent 1 model.*

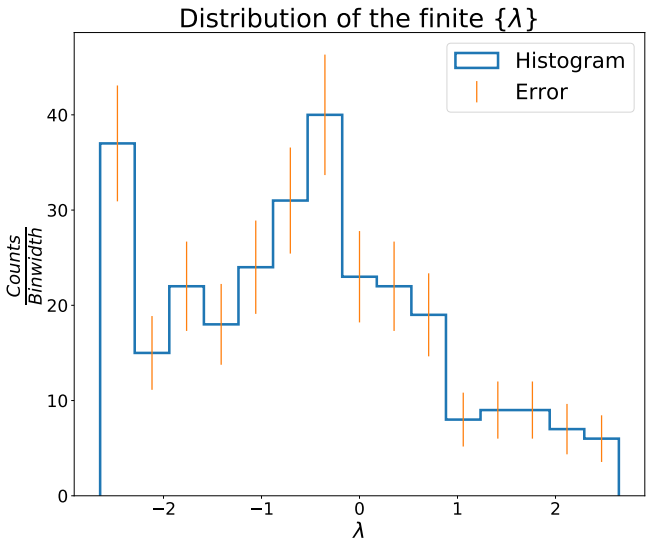

Figure 2: *Distribution of the values of the Lagrange multipliers for the Max Ent 1 model.*

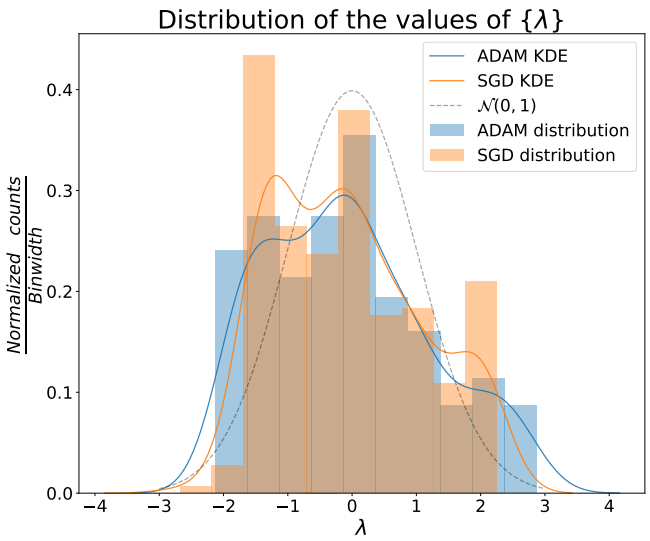

Figure 3: *The two distributions obtained for $\{\lambda\}$ of the Max Ent 2 by using SGD and ADAM, together with a $\mathcal{N}(0,1)$ distribution as reference.*

III. TASK 4: MAX ENT 2

The hamiltonian is now defined as

$$H(\{\sigma\};\{\theta\}) = -\sum_{i=1}^{S} \lambda_i \sigma_i - k \frac{\left(\sum_{i=1}^{S} \sigma_i\right)^2}{S} \qquad (1)$$

where $\{\theta\}$ denotes the family of parameters $\{\{\lambda_i\}_{i=1}^{S}, k\}$. The constraints are

$$\langle 2p_i - 1 \rangle_{emp} = \langle \sigma_i \rangle_{emp} = \langle \sigma_i \rangle_{model} \qquad (2)$$

$$\frac{\langle (S_+ - S_-)^2 \rangle_{emp}}{S} = \frac{\left\langle \left(\sum_{i=1}^{S} \sigma_i\right)^2 \right\rangle_{emp}}{S} = \frac{\left\langle \left(\sum_{i=1}^{S} \sigma_i\right)^2 \right\rangle_{model}}{S}$$

As exact analytical handling is not possible for this model, the suggested course of action was to perform a series of numerical simulations aimed at obtaining approximate values for the set of parameters. The simulation proceeded as follows:

1. initialize the set of parameters $\{\theta\}$. All values have been drawn from a normal distribution, $\theta \sim \mathcal{N}(0,1)$;
2. perform a Metropolis simulation, using the hamiltonian 1;

3. use the last (stabilized) states in the simulation to perform an average of the quantities subject to constraints (2 and III);
4. minimize the Kulbach-Leibler divergence with respect to the set of parameters;
5. adjourn the parameters by means of gradient descent (GD):

$$\theta_j \leftarrow \theta_j + \eta\left(\langle C_j\rangle_{emp} - \langle C_j\rangle_{model}\right), \quad j = 1, ..., S+1$$

6. repeat until convergence.

Two different GD techniques were employed: *Standard Gradient Descent (SGD)* and *ADAM*. Generally speaking, ADAM outperforms SGD. A visualization of the results is provided in fig. 3.

IV. TASK 5: COMPARING MAX ENT 2 AND RANDOM FIELD ISING MODEL

Referring to the analytic treatment of the *random field Ising model (RFIM)*, we identify the external fields h_i with the Lagrange multipliers λ_i. It is worth noting, however, that despite the initialization, the assumption of having independent, gaussian distributed random fields is not included in the numeric solution provided by the Metropolis/gradient descent method employed above. Indeed, fig. 3 seems to exclude any gaussian behaviour.
The RFIM hamiltonian for N connected spins with interaction strength J is:

$$H(\{\sigma\}) = -\sum_{i=1}^{N}\lambda_i\sigma_i - \frac{J}{N}\sum_{i,j}\sigma_i\sigma_j$$

However, the results pertaining to the phase diagram of the RFIM can be adapted to our situation. In particular:

- σ, which was to be intended as the variance of the random fields, is substituted by the variance of the Lagrange multipliers: $\sigma = \sqrt{\mathrm{Var}(\{\lambda\})}$
- the coupling constant J is substituted by the value of k;
- the number of spins N is substituted by the number of species S.

The RFIM has a phase transition curve which follows the following integral curve:

$$2J(\sigma) = \int_{-\infty}^{+\infty}\frac{\mathrm{d}x}{\sqrt{2\pi\sigma^2}}\left(\frac{\exp\left(-\frac{x^2}{2\sigma^2}\right)}{(\cosh x)^2}\right)$$

The hamiltonian of the model at hand is then represented by the point $\left(\frac{\sum_{i=1}^{S}\lambda_i}{S}; k\right)$. The points are rather close to the phase transition line; this may be seen as a hint towards the theory according to which living systems are poised at criticality. However, it must be stressed that the parameters were initialized by drawing them from $\mathcal{N}(0,1)$, but later moved away under the GD procedure, so any comparison with RFIM should be cautious.

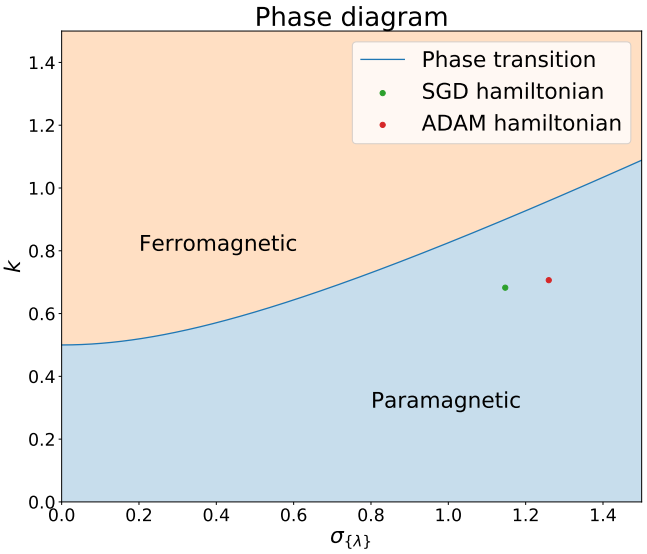

Figure 4: *The phase diagram of the Hamiltonian and the two solutions (SGD and ADAM).*

V. TASK 6: MAX ENT 3

The third approach to Max Ent modelling takes into account the abundances instead of the presences; it also introduces two point correlations. The hamiltonian reads

$$H(\{\vec{x}\}) = -\sum_{i=1}^{S}\lambda_i^{(g)}x_i - \frac{1}{2}\sum_{i,j}^{S}\lambda_{ij}^{(g)}x_jx_i$$
$$\text{where} \quad \lambda_{ij}^{(g)} = \lambda_{ji}^{(g)}$$

so that the parameters are now $\{\lambda^{(g)}, \lambda_{ij}^{(g)}\}$. In the following, $\mathbf{M}$ will be used instead of $\lambda_{ij}^{(g)}$.
The constraints are

$$\langle x_i\rangle_{emp} = \langle x_i\rangle_{model}, \quad \langle x_ix_j\rangle_{emp} = \langle x_ix_j\rangle_{model} \qquad (3)$$

Using the two point correlation function results in $\frac{S(S+1)}{2}$ constraints, as it is symmetric; the abundances, instead, give the usual S constraints.
If we allow the abundances to be treated as continuous variables, the partition function can be written as

$$Z = \int_0^{+\infty} d^sx\exp\left(-\vec{\lambda}\cdot\vec{x} - \vec{x}^T\mathbf{M}\vec{x}\right)$$

We now use the *gaussian approximation*; for each species, the respective integrated term resembles a gaussian integral, the only difference being the integration domain, as $\vec{x} \in {\mathbb{R}_+}^S$. Were we to integrate over $\mathbb{R}^S$, we would perform a usual gaussian integral. To do so, we can argue that, for species with a high enough $\langle x_i\rangle_{emp}$, the portion of gaussian to the left of the 0 mark is negligible (of course, provided a suitable variance), so that $\int_0^{+\infty} \approx \int_{\mathbb{R}^S}$.
In order to employ such an approach, we first had to remove all species which did not meet the "high enough

mean" requirement; we thus only considered the species for which

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

where $\sigma_{x,i}$ is the standard deviation of the i-th species averaged over the N subplots.
We can then calculate Z and, by extension, the various parameters, for the S' remaining species:

$$Z = \int_{\mathbb{R}^{S'}} d^{S'}x \exp\left(-\vec{\lambda}^{(g)} \cdot \vec{x} - \vec{x}^T \mathbf{M} \vec{x}\right) = \frac{(2\pi)^{-S'/2}}{\sqrt{|\mathbf{M}|}} \exp\left(\frac{\vec{\lambda}^{(g)T} \mathbf{M}^{-1} \vec{\lambda}^{(g)}}{2}\right)$$

$$\langle x_i \rangle_{emp} = \langle x_i \rangle_{model} = -\frac{\partial}{\partial \lambda_i^{(g)}} \log Z = \sum_{j=1}^{S'} \mathbf{M}_{ij}^{-1} \lambda_j^{(g)}$$

$$\rightarrow \langle \vec{x} \rangle_{emp} = \mathbf{M}^{-1} \vec{\lambda}^{(g)}$$

$$\frac{\partial^2}{\partial \lambda_i^{(g)} \partial \lambda_j^{(g)}} \log Z = \frac{\partial}{\partial \lambda_j^{(g)}} \langle x_i \rangle_{model} = \langle x_i x_j \rangle_{model} - \langle x_i \rangle_{model} \langle x_j \rangle_{model}$$

$$= \langle x_i x_j \rangle_{emp} - \langle x_i \rangle_{emp} \langle x_j \rangle_{emp} = \mathrm{Cov}(x_i, x_j) = \mathbf{M}_{ij}^{-1}$$

So that, in the end,

$$\vec{\lambda}^{(g)} = \mathbf{M} \langle \vec{x} \rangle_{emp}, \quad \mathbf{M}_{ij}^{-1} = \mathrm{Cov}(x_i, x_j) \quad i, j = 1, ..., S' \tag{4}$$

An additional assumption was included: all the diagonal elements of $\mathbf{M}$ were set to 0, as to avoid self-interaction terms.
Two histograms are depicted (figs 5 and 6) which show the distributions of the obtained parameters. It is worth noting that the vast majority of the $\mathbf{M}_{ij}$ are close to 0, making it a very peaked distribution. From a graph point of view, this makes the associated network very sparse, with few important connections. The values of λ seem instead to follow a multimodal distribution, with a tail to the right.

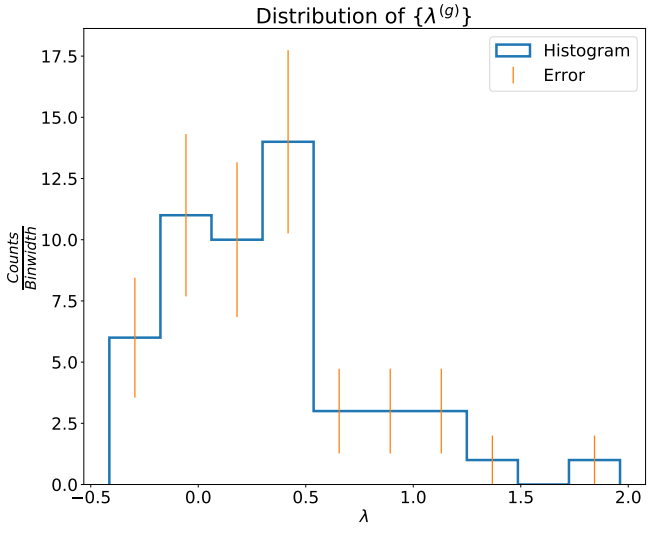

Figure 5: *Histogram of the values of $\vec{\lambda}^{(g)}$.*

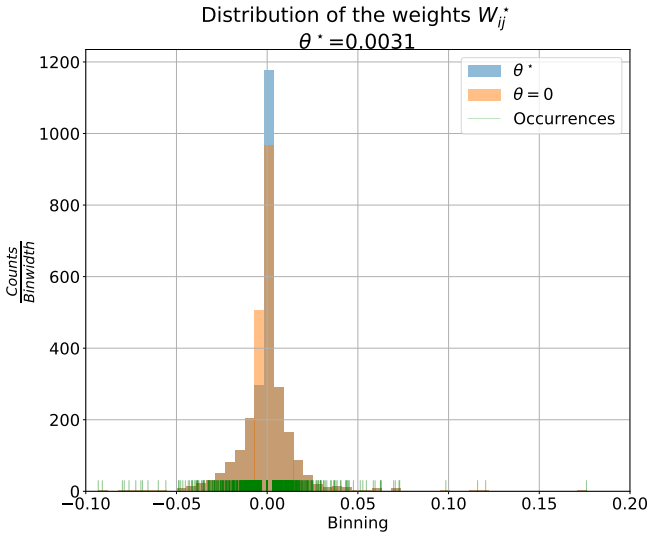

Figure 6: *Histogram of the values of $\mathbf{M}$, before (orange) and after (blue) the trimming. Occurrences after the trimming are highlighted in the lower axis.*

VI. TASK 7: NETWORK PROPERTIES

The matrix $\mathbf{M} = \mathrm{Cov}(\vec{x}, \vec{x})^{-1}$ can be thought of as a two-species interaction matrix, as it appears in the quadratic term in the exponent. From a network point of view, $\mathbf{M}$ is the weighted adjacency matrix. We can then choose a threshold $\theta > 0$ and set $\mathbf{M}_{ij} = 0$ if $|\mathbf{M}_{ij}| < \theta$. This equates to ignoring the weaker interactions, thus leading to networks with less edges; a plot was produced in order to show the relationship between the threshold θ and the number of connected components (fig. 7). Indeed, the next step was to find the value $\theta^{\star}$ above which the network gets split in two components. Its calculation through the eigenvalues of the laplacian yielded $\theta^{\star} = 0.0031$.

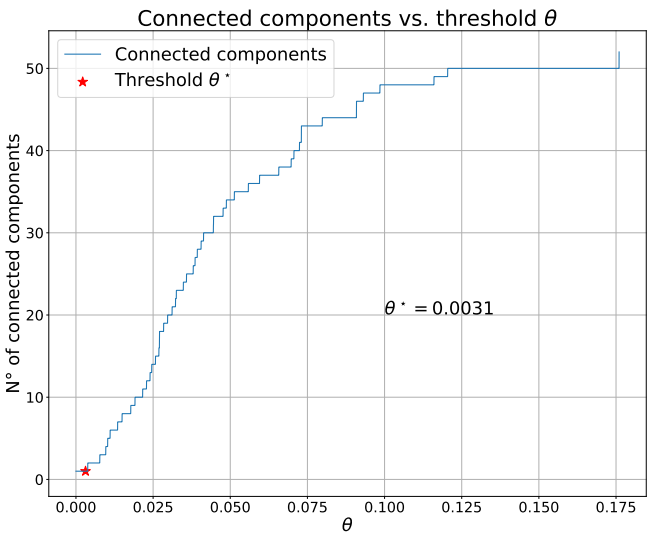

Figure 7: *The number of connected components vs the threshold θ. The value $\theta^{\star}$ is highlighted.*

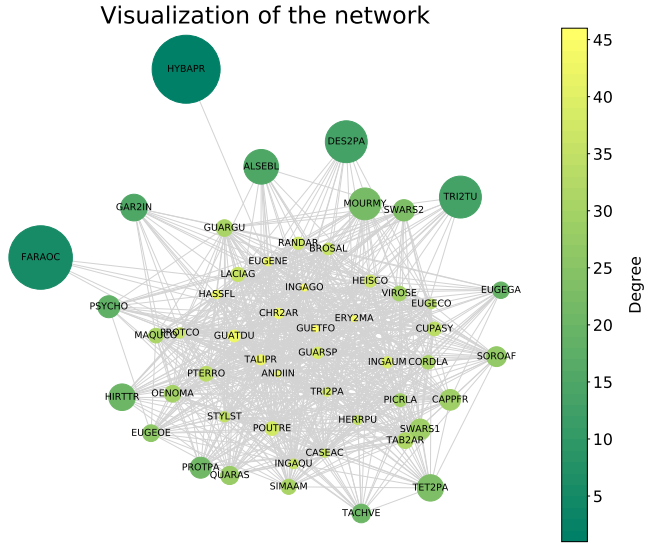

Figure 8: *Visualization of the network $W^\star$; the dimension of a bubble is proportional to the average abundance of that species, while the colouring reflects its number of edges.*

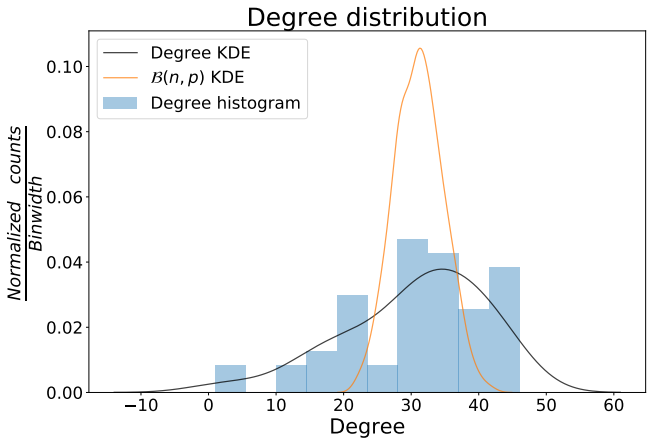

Figure 9: *The degree distribution together with its KDE. The KDE for a set of values drawn from $\mathcal{B}(n,p)$ is depicted as reference.*

The corresponding graph, $W^\star$ (see fig. 8), was then studied and compared to an *Erdős–Rényi (ER) graph*. In particular, some key quantities were compared, namely *diameter, degree distribution, clustering coefficient, degree assortativity coefficient and betweenness centrality.* Given the number of nodes n and the probability of an edge appearing, p, analytical expressions for an ER graph are only available for the degree distribution (which is binomial, $\mathcal{B}(n,p)$) and the clustering coefficient, $C = \frac{c}{n-1} = \frac{\langle k\rangle}{n-1}$. It was then decided to estimate the quantities as averages of an ensemble of $Num = 100$ ER graphs, generated by the Pyhton 3 library *NetworkX* using the number of nodes, n and the probability of an edge, p in order to build an ER graph. Since $W^\star$ had $n_{W^\star} = 52$ and $\langle k\rangle_{W^\star} = 30.42$, we set $n = n_{W^\star} = 52$ and $p = \frac{\langle k\rangle_{W^\star}}{n_{W^\star}-1} \approx 0.597$. All quantities are grouped in table I.

	Forest	ER graph	Compatibility
Diameter	3	2	∞
Assortativity	-0.241	-0.037	9.3
Average degree	30.42	30.40	0.006
Average clustering coefficient	0.729	0.596	9.4
Max betweenness centrality	0.0082	0.0133	4.4

Table I: *Values of the key quantities for both the BCI forest and the average of 100 ER graphs. The ∞ is due to the ER graphs all having diameter 2.*

Even at a glance, $W^\star$ is rather different from the average ER graph.

- **Diameter**: as the value in table I is the result of an average over 100 realizations, the fact that the value is exactly 2 suggests that ER graphs are more connected than $W^\star$.
- **Assortativity coefficient**: as expected, for an ER graph the assortativity is close to 0, as there is no notion of degree-based preference in how the edges are established. For $W^\star$, instead, the assortativity is slightly negative. Visual intuition of this behaviour can be found in fig. 8, as the outermost bubbles, which have lower degree, are not connected to each other.
- **Average degree**: as this parameter was used to build the ER graphs, it is not surprising to see that these values are almost identical (with a compatibility of ≈ 0.006).
- **Average clustering coefficient**: for an ER graph it holds that $C = p$; this quantity is lower than the one calculated for $W^\star$, meaning that on average the BCI network is "more transitive" than a random graph.
- **Maximum betweenness centrality**: for this specific quantity, we decided to take the average of the maximum value for each ER graph, in order to specifically look at the presence of "in-between" nodes. It seems that $W^\star$ has at least one node more central than those in an ER graph.

Another important distinction has to be made on the basis of the degree distribution. Were $W^\star$ similar to an ER graph, one would expect its degree distribution to be approximately binomial (precisely $\mathcal{B}(n,p)$), but as it is shown in fig. 9, that is not the case, as the *kernel density estimate (KDE)* of the degree distribution and of 1000 iid binomial variables (drawn from $\mathcal{B}(n,p)$) are widely different.
In conclusion, $W^\star$ is not similar to an ER graph.

VII. APPENDIX: SADDLE POINT APPROXIMATION

In the Max Ent 2 model, the hamiltonian to be considered is the following:

$$H(\{\sigma\}) = -\sum_{j=1}^{S} \lambda_j \sigma_j - \frac{k}{S}\left(\sum_{j=1}^{S}\sigma_j\right)^2$$

The suggested course of action was to perform a mixture of metropolis algorithm and gradient descent in order to get an approximate value of the parameters for such a model. However, we performed some introductory calculations involving the Hubbard-Stratonovich transformation and the saddle point approximation, as to have a meaningful initialization of the parameters. In order to employ the H.S. transform, we first considered the hamiltonian

$$H(\{\sigma\}) = -\sum_{j=1}^{S} \lambda_j \sigma_j + \frac{k}{S}\left(\sum_{j=1}^{S}\sigma_j\right)^2$$

as using the previous one would give contradictory results. The partition function is

$$\begin{aligned}
Z =& \sum_{\{\sigma\}} \exp(-H(\{\sigma\})) = \\
\xrightarrow{H.S.} \quad & = \sum_{\{\sigma\}} \exp\left[\sum_{j=1}^{S}\lambda_j\sigma_j\right] \int_{-\infty}^{+\infty} \sqrt{\frac{S}{4\pi k}} dx \exp\left[-\frac{Sx^2}{4k} + \left(\sum_{i=1}^{S}\sigma_i\right)x\right] \\
& = \sqrt{\frac{S}{4\pi k}} \int_{-\infty}^{+\infty} dx \exp\left[-\frac{Sx^2}{4k}\right] \prod_{i=1}^{S} \sum_{\sigma_i=\pm 1} \exp\left[\sigma_i\left(x+\lambda_i\right)\right] \\
& = \sqrt{\frac{S}{4\pi k}} \int_{-\infty}^{+\infty} dx \exp\left[-\frac{Sx^2}{4k}\right] \prod_{i=1}^{S} \left[2\cosh(\lambda_i + x)\right] \\
& = \sqrt{\frac{S}{4\pi k}} \int_{-\infty}^{+\infty} dx \exp\left[S\left(-\frac{x^2}{4k} + \frac{1}{S}\sum_{i=1}^{S}\log\left(2\cosh(\lambda_i+x)\right)\right)\right]
\end{aligned}$$

which can be approximated through saddle point:

$$Z \approx Z^* = \sqrt{\frac{S}{4\pi k}}\; e^{S\mathcal{L}(x_m)}, \qquad \mathcal{L}(x) = \left(-\frac{x^2}{4k} + \frac{1}{S}\sum_{i=1}^{S}\log\left(2\cosh(\lambda_i+x)\right)\right)$$

where x_m is to be evaluated through the self-consistency equation

$$x_m: \quad \left.\frac{\partial \mathcal{L}(x)}{\partial x}\right|_{x=x_m} = 0 \quad \rightarrow \quad x_m = \frac{2k}{S}\sum_{i=1}^{S}\tanh\left(x_m + \lambda_i\right)$$

Then, using the saddle point approximation for Z, we can calculate the approximated parameters as before:

$$\frac{\partial \log Z^*}{\partial \lambda_i} = \langle\sigma_i\rangle_{model} \overset{!}{=} \langle\sigma_i\rangle_{emp}$$

$$\frac{\partial \log Z^*}{\partial k} = \frac{\langle\left(S_+ - S_-\right)^2)\rangle_{model}}{S} \overset{!}{=} \frac{\langle\left(S_+ - S_-\right)^2)\rangle_{emp}}{S}$$

which yield, respectively,

$$\langle\sigma_i\rangle_{emp} = \tanh(\lambda_i + x_m), \qquad i = 1, ..., S,$$

and

$$\frac{\langle\left(S_+ - S_-\right)^2)\rangle_{emp}}{S} = -\frac{1}{2k} + \frac{Sx_m^2}{4k^2}$$

These, combined with the self-consistency equation for x_m, give the following expressions for the parameters:

$$\begin{aligned}
&\lambda_i = \tanh^{-1}\left(\langle\sigma_i\rangle_{emp}\right) - x_m, \qquad i = 1, ..., S, \\
&\frac{1}{2k} = -\frac{1}{S}\left[\left\langle\left(\sum_{i=1}^{S}\sigma_i\right)^2\right\rangle_{emp} - \left(\sum_{i=1}^{S}\langle\sigma_i\rangle_{emp}\right)^2\right] = -\frac{\mathrm{Var}(\sum_{i=1}^{S}\sigma_i)_{emp}}{S}
\end{aligned}$$

After some manipulations, we get the following values:

$$\begin{aligned}
x_m &= \frac{2k}{S}\sum_{i=1}^{S}\langle\sigma_i\rangle_{emp}, \\
k &= -\frac{S}{2\mathrm{Var}(\sum_{i=1}^{S}\sigma_i)_{emp}}, \\
\lambda_i &= \tanh^{-1}\left(\langle\sigma_i\rangle_{emp}\right) - x_m, \qquad i = 1, ..., S.
\end{aligned}$$

More in depth comparisons with the results obtained by the suggested methods can be found in the Jupyter Notebook.