Exercise 1 - Application of Max Ent and Network analysis

Deadline: March 31/2019. **Upload in the moodle**:

1. A presentation of the project in .pdf (no more than 6 pages).
2. The source code through which you generated the presented results

The files name should be surname_name_ex1.pdf and surname_name_code1.XX
If the project has been developed by a group (max 3 persons), please indicate the name of the other authors in the presentation of the project (.pdf)

The Dataset: Barro Colorado Forest Census Plot Data (Version 2012) (file: bci05.csv)
The 50-hectare plot at Barro Colorado Island, Panama, is a 1000 meter by 500 meter rectangle of forest inside of which all woody trees and shrubs with stems at least 1 cm in stem diameter have been censused. Every individual tree in the 50 hectares was permanently numbered with an aluminum tag in 1982, and every individual has been revisited six times since (in 1985, 1990, 1995, 2000, 2005, and 2010). The dataset you will analyse is the of year 2005.
In each census, every tree was measured, mapped and identified to species. Details of the census method are presented in Condit (Tropical forest census plots: Methods and results from Barro Colorado Island, Panama and a comparison with other plots; Springer-Verlag, 1998), and a description of the seven-census results in Condit, Chisholm, and Hubbell (Thirty years of forest census at Barro Colorado and the Importance of Immigration in maintaining diversity; PLoS ONE, 7:e49826, 2012).

The file is made by 368123 rows, each one representing an individual tree, and 9 columns denoting:
Individual stem code (tag), species name (sp), x-coordinate (gx), y-coordinate (gy), diameter breast height (dbh),class tag (pom),measurement date (date), type of measurement (codes),status (alive –A or death - D) of the trees (status)

The information you will need to do the exercise are species name and coordinates **for alive** trees only. You can discard all other information.

Tasks

1. Extract the information you need from the raw data. How many species S there are in sampled plot?

2. Divide the 50-hectar plots in subplot of 0.25 hectare each. We will assume that these N=200 plots are independent and we can do the statistics on these plots. Calculate:
   a) the vector of the abundances for all the species ($x_1,\dots,x_S$) for each subplot;
   b) the average "presence" p_i of each species (i=1,,S) averaged over the N subplots.

3. Build a Max Ent model 1 with constraints only given by $(1+ m_i)/2= p_i$, with $\langle\sigma_i\rangle_{emp} = m_i$. Which are the Lagrangian parameters that satisfies the constraints? Discuss the result.
   Suggestion: You do not need to simulate anything

4. Build a Max Ent model 2 with Hamiltonian $H=-\sum_{j=1}^{S}\lambda_j\sigma_j-\frac{k}{N}\left(\sum_{j=1}^{S}\sigma_j\right)^2$ with constraints $(1+ m_i)/2 = p$ (i=1,….,S) and $\langle\left(\sum_{j=1}^{S}\sigma_j\right)^2\rangle_{emp}=\langle S_+ - S_-\rangle_{emp}$, where S_+ and S_- are

respectively the number of species present and absent (in a plot), and then you need to take the average over all plots.

5. *Optional. H* is the Hamiltonian of the Random Field Model [see Schneider and E. Pytte paper (PRB 1977) and the class Models of Theoretical Physics]. Check if the average of λ_j is compatible with 0. If so, show where the variance of λ_j and the k/N inferred at point 4 are in the phase diagram of the Hamiltonian *H.*

6. Taking in account information on the populations of the most abundant species in each plot, build a Max Ent model 3, using as constraints the average abundance $\langle x_i \rangle_{emp} = \langle x_i \rangle_{model}$ and the two-point correlation function $\langle x_i x_j \rangle_{emp} = \langle x_i x_j \rangle_{model}$ and estimate the Lagrangian parameters λ_j and M_{ij} (with $M_{ij} = M_{ji}$ and $M_{ii}=0$) using the Gaussian approximation. In order to evaluate the most abundant species, consider only those species for which $\langle x_i \rangle_{emp} - 2 * \sigma_{x,i} > 0$, where $\sigma_{x,i}$is the standard deviation of the species *i* calculated from the data (over the subplots).

7. What distributions do the entries of M_{ij} follow? Select a threshold θ and put to zero all interactions smaller than this threshold, i.e. $M_{ij} = 0$ if$| M_{ij} | < \theta$. The trimmed matrix can be considered the weighted adjacency graph W of the species interactions networks. *a)* Calculate the number of connected components in the graph as a function of θ; *b*) Consider the graph W^* with the threshold θ^* so that for $\theta > \theta^*$ the graph is not made by one single connected. In other words, θ^* is the largest threshold for which the graph is connected. *c*) Analyze the structural properties of W^* (degree distribution, diameter, clustering, degree assortativity, betweeness centrality). In what respect, if any the graph is different from a random ER graph?