Advanced Probabilistic Machine Learning and Applications

Diego Baptista and Caterina De Bacco

December 5, 2021

Tutorial 8: The Stochastic Block Model and the degree-corrected SBM

Exercise 1: implementing various inferences for the standard SBM

In this tutorial we will implement various inference techniques and models to solve the SBM on real networks. We will use several codes developed in the package *pysbm* that can be found at https://github.com/funket/pysbm. This python module contains several objective functions and inference procedures, including some of those seen in Lecture 8.

(a) Clone the github repository **pysbm**.

(b) Download the datasets of **American College football** *(football)*, **Zachary's karate club** *(karate)* and **Political blogs** *(polblogs)* from http://www-personal.umich.edu/~mejn/netdata/ and put them inside the folder *pysbm/Network Data/*.

(c) Run three different inference procedures using the weighted SBM, i.e. the model with the Poisson likelihood. We suggest to run the greedy algorithm proposed by Karrer and Newman (2011) and two versions of a Monte Carlo Metropolis-Hasting scheme proposed by Peixoto (2014). Comment on their differences.

(d) Plot the adjacency matrices ordered by the inferred blocks and compare with the unordered one.

(e) Plot the affinity matrices of two partitions at your choice.

Exercise 2: degree-corrected SBM (DC-SBM)

As you could notice in the previous exercise, the best partition found by the algorithms favours a block division correlated with degree.
In fact, maximizing the KL divergence between the SBM probability and a random uniform null model $p_0(r,s)$ as the one seen in the Lecture 8 encourages the optimal blocks to be correlated to the degree of a node, which is quite unrealistic. In other words, blocks are made of nodes of similar degree. The solution to this problem is to incorporate explicitly degree heterogeneity into the model as in the so called *degree-corrected* SBM introduced in Karrer and Newman (2011). This implies introducing new hidden variables $\theta_i \in \mathbb{R} \geq 0$ controlling the expected degree of node i. It works as follows:

$$
\begin{aligned}
P(\mathbf{A}|\theta, q, C) &= \prod_{i<j} \text{Pois}\left(A_{ij}; \theta_i \theta_j\, C_{q_i q_j}\right) && (1)\\
&= \prod_{i<j} \frac{e^{-\theta_i \theta_j\, C_{q_i q_j}}\ \left(\theta_i \theta_j\, C_{q_i q_j}\right)^{A_{ij}}}{A_{ij}!} \quad . && (2)
\end{aligned}
$$

One can normalize this new parameter as:

$$
\sum_i \theta_i \delta_{q_i, r} = 1 \quad \forall r = 1, \dots, K \quad . \qquad (3)
$$

Then θ_i can be interpreted as the probability that an edge connected to the group q_i lands to i itself.

(a) Derive the null model suited for the KL divergence representation of the DC-SBM as done in the Lecture 8 for the standard SBM.
Comment on it.

(b) Run the same inference as before, but this time using the degree-corrected likelihood as objective function.
Comment on the different partitions obtained compared to the standard SBM.

Exercise 3

Choose two inference methods and apply similar analysis for the football network (K=11) and the political blogs one (K=2).

References

B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).

T. P. Peixoto, Phys. Rev. E **89**, 012804 (2014).