L1: stochastic block model

Caterina De Bacco

April 10, 2025

1. The stochastic block model: the problem

The stochastic block model (SBM) is a model aimed at clustering network data, or community detection on networks.

Imagine you have a dataset of pairwise interactions between individuals. For instance, consider a social network where people interact through friendship or other social relationship. This dataset can be represented using a so called *adjacency* matrix $\mathbf{A}$ such that the entries $A_{ij} = 1$ if person i is a friend of person j and 0 otherwise. This can be represented as a network where people are nodes and there is an edge between two people if they interact (i.e. if they are friends in this example). The network is represented as $G(V, E)$, where V is the set of nodes and E the set of edges. Typically, the number of nodes is denoted as $N = |V|$; an edge is denoted as (i, j). See Figure 1 for an example of such network.

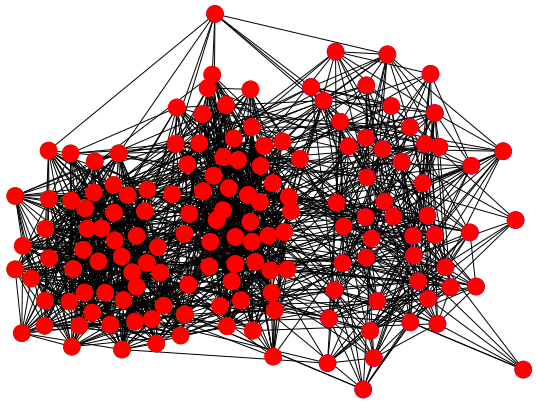

FIGURE 1. Example of a SBM-generated graph.

Objective: cluster people based on their pattern of interaction.

REMARK 1. *The notion of similarity is arbitrary based on the application. Here we assume that two people are similar if they interact in a similar way.*

1.1. Stochastic equivalence

An approach to address this problem is to assume that people interact based on their membership to a particular community. For instance, people that play a particular sport are more likely to interact with people that also play that sport. This pattern is called *assortativity*. The opposite scenario, where two people belonging to *different* community are more likely to interact, is called *disassortativity*. See Figure 2 for an example of these cases. Other types of topologies are also possible.

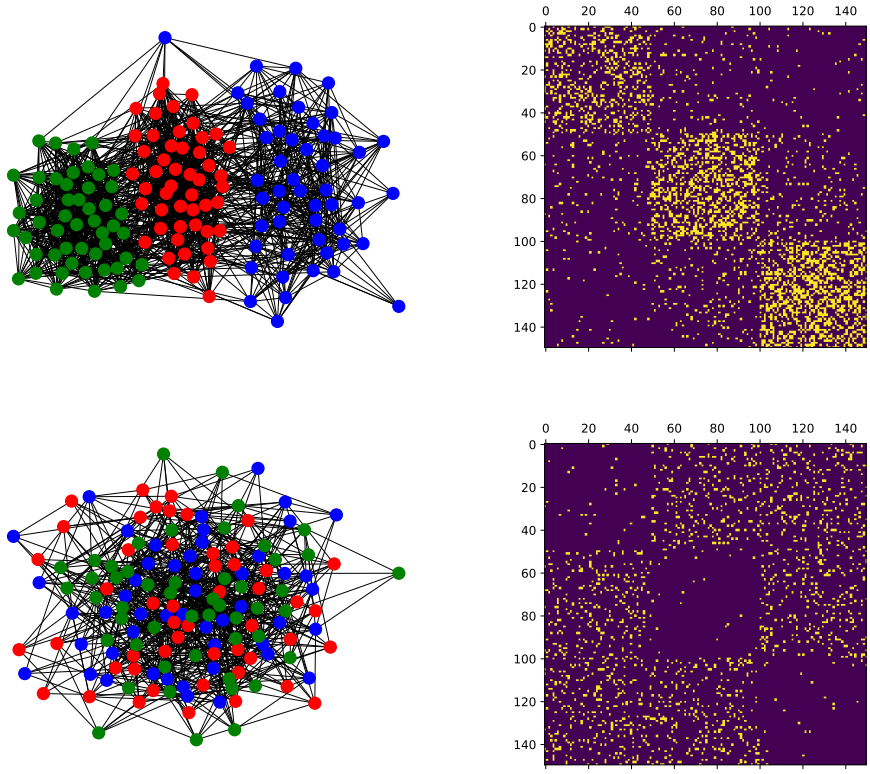

FIGURE 2. Example of a SBM-generated graph. Top) Assortative structure, this is also the one corresponding to Figure 1. Bottom) Disassortative structure. On the right we have the adjacency matrix.

Regardless the structure type, this problem can be formalized by introducing a set of hidden variables, the memberships:

$$q_i \in \{1, \ldots, K\} \quad \text{for} \quad i = 1, \ldots, N \quad , \tag{1}$$

where K is the number of groups/blocks/communities. Now, given two people i and j, we can make the assumption that the probability of them interacting, depends only upon their membership q_i and q_j:

$$Pr\left((i, j) \in E\right) = f\left(q_i, q_j\right) \quad . \tag{2}$$

This is the concept of *stochastic equivalence*. Under this notion, if i and k belong to the same community q, they have the same probability of interacting with a third person j. In

other words, the pattern of interactions is completely captured by the hidden variables, no other microscopic details about the single people matter.
All we need to know to model the network based on the observed data, is the set of memberships and the interaction pattern, i.e. how interactions happen between members of the same or different groups. This is regulated formally by a $K \times K$ matrix $\mathbf{C}$ called *affinity* matrix where the entries C_{kq} are proportional to the probability that a member of group k interacts with a member of the group q.
The SBM is a probabilistic model that considers two set of parameters, the membership $\mathbf{q} = (q_1, \ldots, q_N)$ and the affinity matrix $\mathbf{C}$, and makes the stochastic equivalence assumption.

Objective: estimating the parameters $\theta := (\mathbf{q}, \mathbf{C})$, the membership and affinity matrix, given the observed interactions, the adjacency matrix $\mathbf{A}$, i.e. performing inference.
Once we have the membership, we can cluster people based on where they belong to, which was our original goal.

REMARK 2. *Thanks to the group assignments and the affinity matrix, the model is flexible as it can represent different structures.*

REMARK 3. *It is analytically tractable and computationally scalable. Several inference methods can be applied.*

1.2. The standard model (binary entries)

Given the group assignments $\mathbf{q}$, the edges are generated independently and the probability of their existence is:

$$Pr\left((i,j) \in E\right) = C_{q_i q_j} \quad . \tag{3}$$

For simplicity, in all the following, we do not allow self-edges, i.e. $A_{ii} = 0$, and consider undirected network, i.e. symmetric $\mathbf{A}$. Formally, we can define a likelihood function for the data considering a Bernoulli distribution as:

$$P(\mathbf{A}|\theta) = \prod_{i<j} \text{Bern}\left(A_{ij}; C_{q_i q_j}\right) \tag{4}$$

$$= \prod_{i<j} C_{q_i q_j}^{A_{ij}} \left(1 - C_{q_i q_j}\right)^{(1-A_{ij})} \quad . \tag{5}$$

1.3. Inference: Maximum Likelihood approach

If we assume no prior knowledge about the group assignments, i.e. they are distributed uniformly at random, we can find the parameters by maximizing the likelihood (5). By defining n_r the number of nodes in community r, such that $N_{rs} = \delta_{rs}\binom{n_r}{2} + (1 - \delta_{rs})\, n_r\, n_s$ is the total number of possible edges between group r and s, we can rewrite the likelihood changing the product series to run over all pairs of groups, rather than over all pairs of vertices, thanks to the stochastic equivalence:

$$P(\mathbf{A}|\theta) = \prod_{r>s} C_{rs}^{m_{rs}} \left(1 - C_{rs}\right)^{N_{rs} - m_{rs}} \quad , \tag{6}$$

where m_{rs} is the number of edges between group r and s.
Taking the derivative w.r.t C_{rs} we get:

$$\frac{\partial P(\mathbf{A}|\theta)}{\partial C_{rs}} = C_{rs}^{m_{rs}} (1 - C_{rs})^{N_{rs} - m_{rs}} \left[\frac{m_{rs}}{C_{rs}} - \frac{N_{rs} - m_{rs}}{1 - C_{rs}} \right] \quad , \tag{7}$$

setting it to zero gives:

$$\hat{C}_{rs} = \frac{m_{rs}}{N_{rs}} \quad . \tag{8}$$

We can substitute it inside (6) to get:

$$P(\mathbf{A}|\theta) = \prod_{r,s} \left(\frac{m_{rs}}{N_{rs}} \right)^{m_{rs}} \left(1 - \frac{m_{rs}}{N_{rs}} \right)^{N_{rs} - m_{rs}} \quad . \tag{9}$$

Taking the logarithm yields:

$$\mathcal{L}(\mathbf{A}|\theta) := \log P(\mathbf{A}|\theta) = \sum_{r,s} \left[m_{rs} \log \left(\frac{m_{rs}}{N_{rs}} \right) + (N_{rs} - m_{rs}) \log \left(1 - \frac{m_{rs}}{N_{rs}} \right) \right] \quad . \tag{10}$$

Applying the rules of logarithm and collecting terms yields:

$$\mathcal{L}(\mathbf{A}|\theta) = \sum_{r,s} \left[m_{rs} \log m_{rs} + (N_{rs} - m_{rs}) \log (N_{rs} - m_{rs}) - N_{rs} \log N_{rs} \right] \quad , \tag{11}$$

which is a function only of the counts induced by $\mathbf{q}$. We further have the constraints that:

$$\sum_{q=1,\ldots,Q} n_q = N \tag{12}$$

$$\sum_{r<s=1,\ldots,Q} m_{rs} = |E| \quad , \tag{13}$$

again assuming undirected graphs, i.e symmetric $\mathbf{A}$.

Example. See for instance the network of Figure 3. Using Eq. (9), the partition on the left has $P(\mathbf{A}|\theta) = 0.0433$ whereas the one on the right has $P(\mathbf{A}|\theta) = 0.0002$, i.e. the one on the left has much higher likelihood than the other.

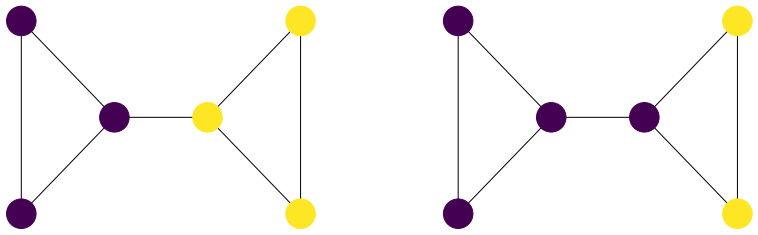

FIGURE 3. Example of network with two triangles. Left) $\mathcal{L} = -3.139$; right: $\mathcal{L} = -8.318$.

Inference of group memberships can be performed by a greedy algorithm based on performing Monte Carlo moves described in Karrer and Newman (2011) as follows:

a. Initialize the system with groups assignments $\mathbf{q}$ extracted uniformly at random.
b. Select a candidate node i and assign a new group q_i^{new} based on the group q that maximizes the change in log likelihood $\Delta L = \log P(\mathbf{A}|\theta^{new}) - \log P(\mathbf{A}|\theta^{old})$. Notice that it can also be negative, i.e. there can be moves that decrease the likelihood.
c. Repeat for all nodes.
d. Take the state that had highest log likelihood and set this as the starting point for a new iteration.
e. Repeat until the log likelihood does not increase anymore (convergence).

Another approach is to use MCMC Metropolis-Hasting methods as we saw in previous tutorials. The simplest implementation of this protocol for the inference of SBMs is to start from a random partition $\mathbf{q}_0$, and use move proposals where a node i is randomly selected, and then its new group membership q_i' is chosen randomly between all $K+1$ choices (where the remaining choice means we populate a new group) according to this probability:

$$P(q_i'|\mathbf{q}) = \frac{1}{K+1} \quad . \tag{14}$$

One then has to decide whether to make the move or not, based on the probability ratio $P(\mathbf{q}'|\mathbf{A})/P(\mathbf{q}|\mathbf{A})$, which determines how much "it costs" to make the move. Calculating $P(\mathbf{q}'|\mathbf{A})/P(\mathbf{q}|\mathbf{A})$ is usually efficient (we do not need to calculate normalization constants, they cancel out).
Another approach is to use variational methods like Variational Expectation Maximization (VEM). We will cover this in the next Lectures.

1.4. Weighted network

If the network is weighted, so that $\mathbf{A}$ is not binary anymore but can have other positive and discrete values, it is better to consider a Poisson distribution for the likelihood:

$$P(\mathbf{A}|\theta) = \prod_{i<j} \text{Pois}\left(A_{ij}; C_{q_i q_j}\right) \tag{15}$$

$$= \prod_{i<j} \frac{e^{-C_{q_i q_j}} C_{q_i q_j}^{A_{ij}}}{A_{ij}!} \quad . \tag{16}$$

Following similar reasonings to before, i.e. using the stochastic equivalence, we get:

$$P(\mathbf{A}|\theta) = \prod_{r<s} \frac{e^{-N_{rs} C_{rs}} C_{rs}^{m_{rs}}}{A_{ij}!} \quad . \tag{17}$$

Proceeding again in an MLE fashion, we get by simple differentiation:

$$\hat{C}_{rs} = \frac{m_{rs}}{N_{rs}} \quad , \tag{18}$$

which is the same as for the Bernoulli (binary) case. Substituting back into the likelihood, taking the log and ignoring constants we get the (unnormalized) loglikelihood:

$$\log P(\mathbf{A}|\mathbf{q}, \hat{C}) = \sum_{rs} m_{rs} \log \frac{m_{rs}}{N_{rs}} \quad . \tag{19}$$

1.4.1. MLE interpretation

Let's now rewrite that expression in a different way by applying some simple algebra:

$$\log P(\mathbf{A}|\mathbf{q}, \hat{C}) = \sum_{rs} \frac{m_{rs}}{2|E|} \log \frac{m_{rs}/2|E|}{N_{rs}/N^2} \quad , \tag{20}$$

where we have neglected irrelevant constants. If we now pick one edge uniformly at random, the probability that its end nodes have group assignment r and s is $p_K(r,s) = \frac{m_{rs}}{2|E|}$. If the network had the same group assignment but edges where assigned completely at random, i.e. with no dependence on the group membership, this probability would be equal to N_{rs}/N^2 instead. This is the denominator of the logarithm inside Eq. (20). We call this second distribution $p_0(r,s)$. We can then rewrite:

$$\log P(\mathbf{A}|\mathbf{q}, \hat{C}) = \sum_{rs} p_K(r,s) \log \frac{p_K(r,s)}{p_0(r,s)} \quad , \tag{21}$$

which is the KL divergence between the two distributions. We can conclude that the most likely group assignments under the ordinary stochastic blockmodel are then those assignments that require the most information to describe starting from a model that does not have group structure.

1.5. Choosing the number of groups K

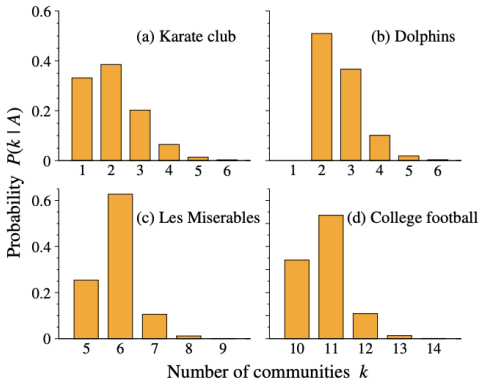

FIGURE 4. Example of finding the best K. This approach calculates the posterior $P(K|\mathbf{A})$ using a Bayesian approach. Figure taken from Newman and Reinert (2016).

In all this lecture we assume that K is given or known.
In real situation this is not the case and we should write explicitly K as an additional parameter of the model, i.e. $\theta := (\mathbf{q}, \mathbf{C}, K)$. However, its inference is usually different than that of the other hidden variables. One in fact treats the problem of choosing the best K as a *model selection* problem in and of itself.
Popular ways to perform model selection are calculating metrics like AIC, BIC or minimum description length. These calculate the cost of the amount of information needed compare to the likelihood score.
Another popular approach is to perform cross-validation: split the dataset into train and

test sets. Run inference on the train and use the parameters obtained there to calculate performance metric in the test set. Repeat for different values of K and choose the one with best performance on the test.
A different approach is to consider a Bayesian formalism with hyper-parameters that depend on K explicitly, and thus calculate the posterior $P(K|\mathbf{A})$ by marginalizing out all the remaining parameters from the full posterior, see for instance Newman and Reinert (2016) or Peixoto (2017) for more details.

1.6. SBM: summary

- SBM is a classical inference problem that allows flexibility to model various datasets and systems.
- It's a nice testbed to try different inference techniques as discussed here.
- It can be further generalized to include more insights specific to the application considered, like dynamicity, overlapping communities, multilayer networks etc...
- It raises several interesting and deep questions like what to do with the final partitions, should we sample or optimize the division in group? How do we select for number of groups?
- As for other similar problems, it is affected by the detectability problem, i.e. there is a rang of parameters within which there is no chance to recover the ground truth partition.

Two references for this lecture are Funke and Becker (2019); Peixoto (2017).

References

M. E. Newman and G. Reinert, Physical review letters **117**, 078301 (2016).
T. P. Peixoto, arXiv preprint arXiv:1705.10225 (2017).
B. Karrer and M. E. Newman, Physical review E **83**, 016107 (2011).
T. Funke and T. Becker, PloS one **14**, e0215296 (2019).