1693

Assortative-Constrained Stochastic Block Models

Daniel Gribel, Thibaut Vidal, Michel Gendreau

ICPR 2020

Abstract

Stochastic block models (SBMs) are often used to find *assortative* community structures in networks, such that the probability of connections within communities is higher than in between communities. However, classic SBMs are not limited to assortative structures. In this study, we discuss the implications of this model-inherent indifference towards assortativity or disassortativity, and show that this characteristic can lead to undesirable outcomes for networks which are presupposedy assortative but which contain a reduced amount of information. To circumvent this issue, we introduce a constrained SBM that imposes strong assortativity constraints, along with efficient algorithmic approaches to solve it. These constraints significantly boost community recovery capabilities in regimes that are close to the information-theoretic threshold. They also permit to identify structurally-different communities in networks representing cerebral-cortex activity regions.

Background

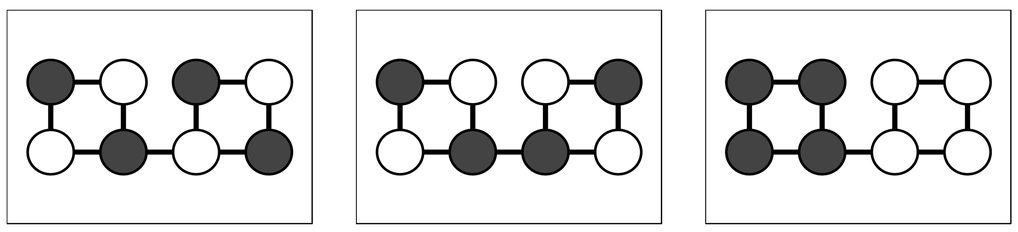
STOCHASTIC BLOCK MODELS. Fitting the parameters of a stochastic block model (SBM) to a given graph is a prominent way of searching for communities. The degree-corrected SBM (DC-SBM), in particular, allows non-uniform node degree distributions, making block modeling more representative of real-world networks:

$$\log P(A|\Omega,Z) = \frac{1}{2}\sum_{rs}^{K}\sum_{ij}^{N}\left(A_{ij}\log(\omega_{rs}) - \frac{k_i k_j}{2m}\omega_{rs}\right)z_{ir}z_{js},$$

in which k_i is the degree of node i, variables Z represent the binary community assignments, and Ω is a symmetric K x K edge probability matrix.

SBM-based community detection approaches, however, are agnostic to the assortativity of their solutions. They can indifferently model assortative and disassortative structures.

This modeling capability can be viewed as an asset but also as a weakness. In the most dramatic situations, non-assortative solutions might go under the radar and lead to mistakes of interpretation. In other cases, non-assortative solutions with a better likelihood may substitute the assortative solutions which were originally sought. This later situation is especially prevalent in case studies involving sparse graphs, or with lightly assortative structures which challenge detection algorithms.



(a) Opt. solution log L = -2.7616

(b) 2nd best solution log L = -5.3193

(c) 3rd best solution log L = -5.9012

Figure 1: The three best solutions in a small example case

CONTRIBUTIONS. In this work, we propose a variant of the DC-SBM which includes user knowledge about assortativity. We incorporate this information by setting assortativity constraints on the DC-SBM parameter set. The key contributions of this work are the following:

1. We introduce a DC-SBM variant which incorporates assortativity constraints to represent prior user knowledge;
2. We propose an efficient solution approach based on local optimization and interior-point algorithms for this model;
3. Through extensive computational experiments, we discuss the practical implications of this constrained model and identify the regimes in which it contributes to improve community detection practice.

Assortative-Constrained SBM

ASSORTATIVITY CONSTRAINTS. Following [Amini and Levina, 2018], two main notions of assortativity can be distinguished for block models:

Strong assortativity. All diagonal terms of the SBM matrix are greater or equal than all off-diagonal terms:

$$\omega_{qq} \geq \omega_{rs} \quad \forall\, q,r,s \in \{1,\ldots,K\}, r \neq s.$$

Weak assortativity. Each diagonal term of the SBM matrix is greater or equal than the other terms in its row:

$$\omega_{qq} \geq \omega_{qs} \quad \forall\, q,s \in \{1,\ldots,K\}.$$

In this study, we will use the strongest definition of assortativity based on the first condition above. With these constraints, the log-likelihood maximization model becomes:

$$\begin{aligned}
\max_{\Omega,Z,\lambda} \quad & \frac{1}{2}\sum_{rs}^{K}\sum_{ij}^{N}\left(A_{ij}\log(\omega_{rs}) - \frac{k_i k_j}{2m}\omega_{rs}\right)z_{ir}z_{js} \\
\text{s.t.} \quad & \omega_{qq} \geq \lambda \quad \forall q \in \{1,\ldots,K\} \\
& \omega_{rs} \leq \lambda \quad \forall r,s \in \{1,\ldots,K\}, r \neq s \\
& \omega_{rs} \geq 0 \quad \forall r,s \in \{1,\ldots,K\},
\end{aligned}$$

where λ represents a continuous variable acting as a threshold.

Methodology

We introduce an iterative algorithm to solve the presented model above. This algorithm starts with a random initial solution and proceeds by iteratively evaluating each possible relocation of a node to a different community. Each such relocation is only applied if its application combined with an optimal update of the SBM matrix results into an improvement of the likelihood. As such, the evaluation of each relocation may require the solution of a small constrained convex optimization subproblem with K^2 variables and constraints to find an optimal SBM matrix for the new partition. Thus, the algorithm combines two techniques:

- **an incremental move evaluation approach**, using the log-likelihood of the unconstrained problem to filter relocation candidates, and possibly keeping this solution if it naturally satisfies the assortativity constraints;
- **an efficient interior point solver** for the following convex constrained subproblem, only used if the relocation candidate was not filtered out due to the previous conditions:

$$\begin{aligned}
\max_{\Omega,\lambda} \quad & \frac{1}{2}\sum_{rs}^{K}\left(m_{rs}\log(\omega_{rs}) - T_{rs}\omega_{rs}\right) \\
\text{s.t.} \quad & \omega_{qq} \geq \lambda \quad \forall q \in \{1,\ldots,K\} \\
& \omega_{rs} \leq \lambda \quad \forall r,s \in \{1,\ldots,K\}, r \neq s \\
& \omega_{rs} \geq 0 \quad \forall r,s \in \{1,\ldots,K\},
\end{aligned}$$

where m_{rs} represents the number of edges between communities r and s according to the fixed partition and

$$T_{rs} = \left(\sum_{t}^{K} m_{rt}\sum_{t}^{K} m_{st}\right)/2m$$

1693

Assortative-Constrained Stochastic Block Models

Daniel Gribel, Thibaut Vidal, Michel Gendreau

ICPR 2020

Experimental Results

SYNTHETIC NETWORKS. To compare the results of the DC-SBM and AC-DC-SBM, we generate 50 synthetic data sets with N = 100 nodes and K = 4 blocks. For each data set, the SBM parameters are uniformly sampled in the following intervals:

$$\begin{aligned} \omega_{rr} &\in [0.45, 0.55] \qquad \forall r \in \{1, \ldots, K\} \\ \omega_{rs} &\in [0, 0.4] \qquad \forall r, s \in \{1, \ldots, K\}, r \neq s. \end{aligned}$$

Figure 2 compares the NMI obtained with the standard DC-SBM and the proposed AC-DC-SBM on these networks. For each network and model, we conduct 50 independent runs from different initial solutions. AC-DC-SBM obtains on 49 out of 50 datasets a better or equal median NMI than DC-SBM. DC-SBM appears to be very sensible to low-quality local minima, and this behavior is particularly visible on the first six data sets presented in the figure. A pairwise Wilcoxon test comparing the average NMI of both methods over the 50 data sets confirms the statistical significance of this difference of performance (with $p = 3.9 \times 10^{-10}$).

Figure 3 compares the number of assortative communities found by AC-DC-SBM and DC-SBM. The standard DC-SBM produces much fewer assortative communities in average (2.43 compared to 3.76). With the AC-DC-SBM, non-assortative partitions are heavily penalized from a likelihood perspective and therefore generally avoided.

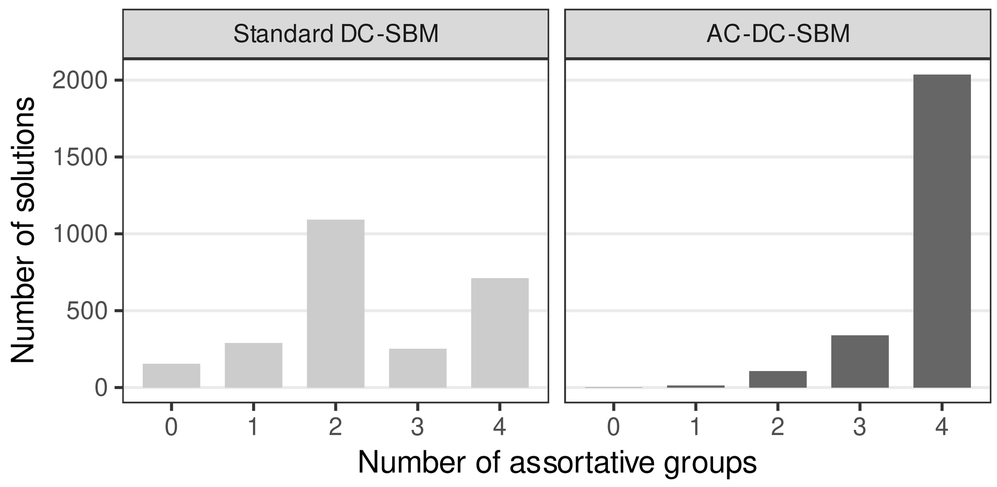


Figure 3: Distribution of the number of assortative communities found by AC-DC-SBM and DC-SBM.

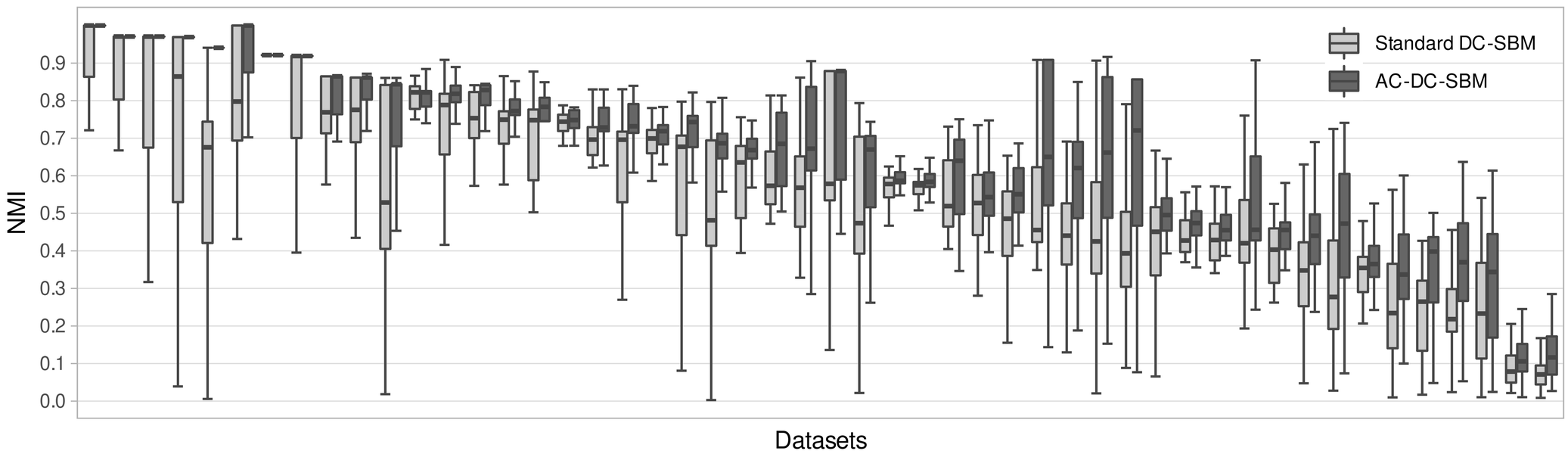
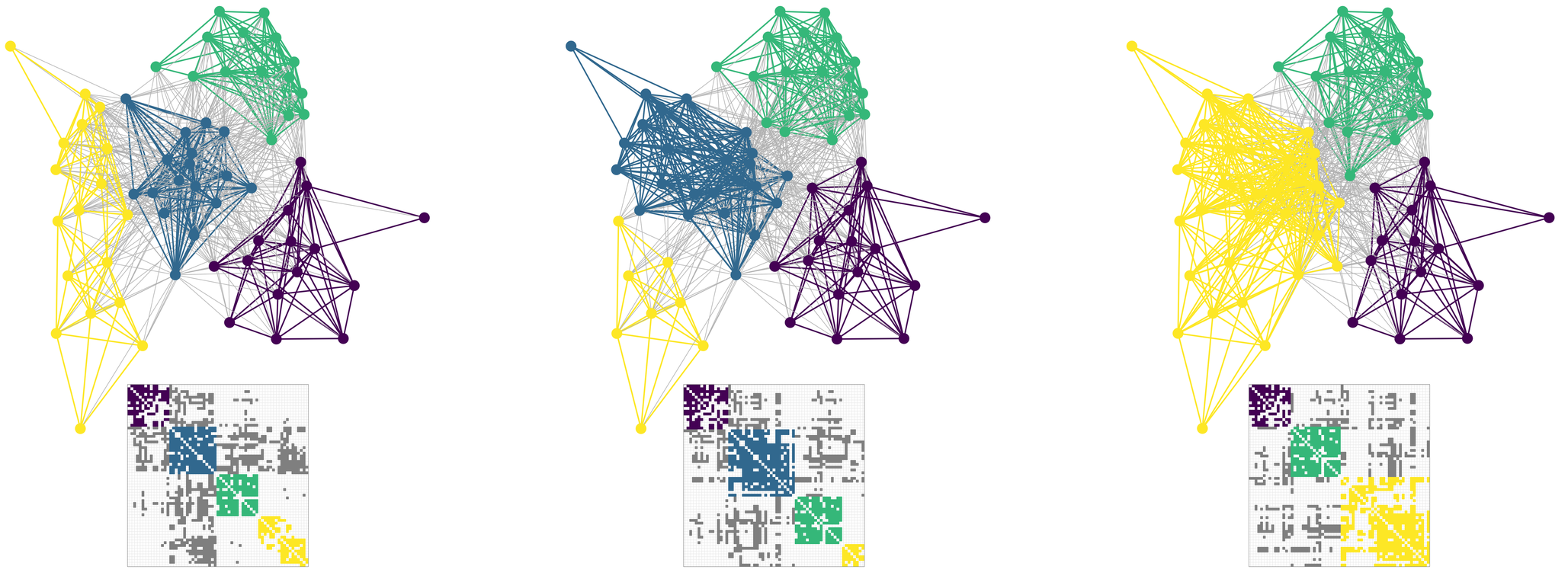


Figure 2: Performance of DC-SBM and AC-DC-SBM on synthetic SBM networks. The results are ordered by median NMI.

BRAIN CORTEX NETWORKS. We analyze in this section the case of the "cats cortex network", which is known to have an assortative structure and is divided into four main functional areas: visual, auditory, frontolimbic, and somatosensory-motor duties. The network is obtained from a connectivity pattern based on 1139 cortico-cortical connections and 65 cortical areas. As in most community detection tasks, the ground truth in this network is not available. In fact, there is no unique "correct" partitioning, but different algorithms can allow to highlight different underlying structures. Figure 4 reports the communities found with the standard DC-SBM, the AC-DC-SBM and modularity maximization models on this dataset.

The best solution obtained with the standard DC-SBM is visibly non-assortative. The minimum value found along the diagonal of the SBM matrix is 1.5060, whereas the maximum value in the off-diagonal is 1.9050. In contrast, the partition produced by the AC-DC-SBM satisfies the strong assortativity conditions. The minimum value of the diagonal of the SBM matrix is 2.0196, and the maximum value in the off-diagonal is 1.7152. Finally, the modularity-maximization approach leads to the most assortative partitioning of this network. Yet, since the model does not take K into consideration, this partitioning contains only three groups, contrasting with the four functional areas which were originally expected.



(a) Standard DC-SBM **(b) AC-DC-SBM** **(c) Modularity maximization model**

Figure 4: The best among 100 network partitions found by different models in the cats cortex network.