Community detection - Bayesian inference for robust detection of assortative structure

Lizhi Zhang [1] Tiago P. Peixoto [1 2]

[1]Department of Mathematical Science, University of Bath

[2]Department of Network and Data Science, Central European University

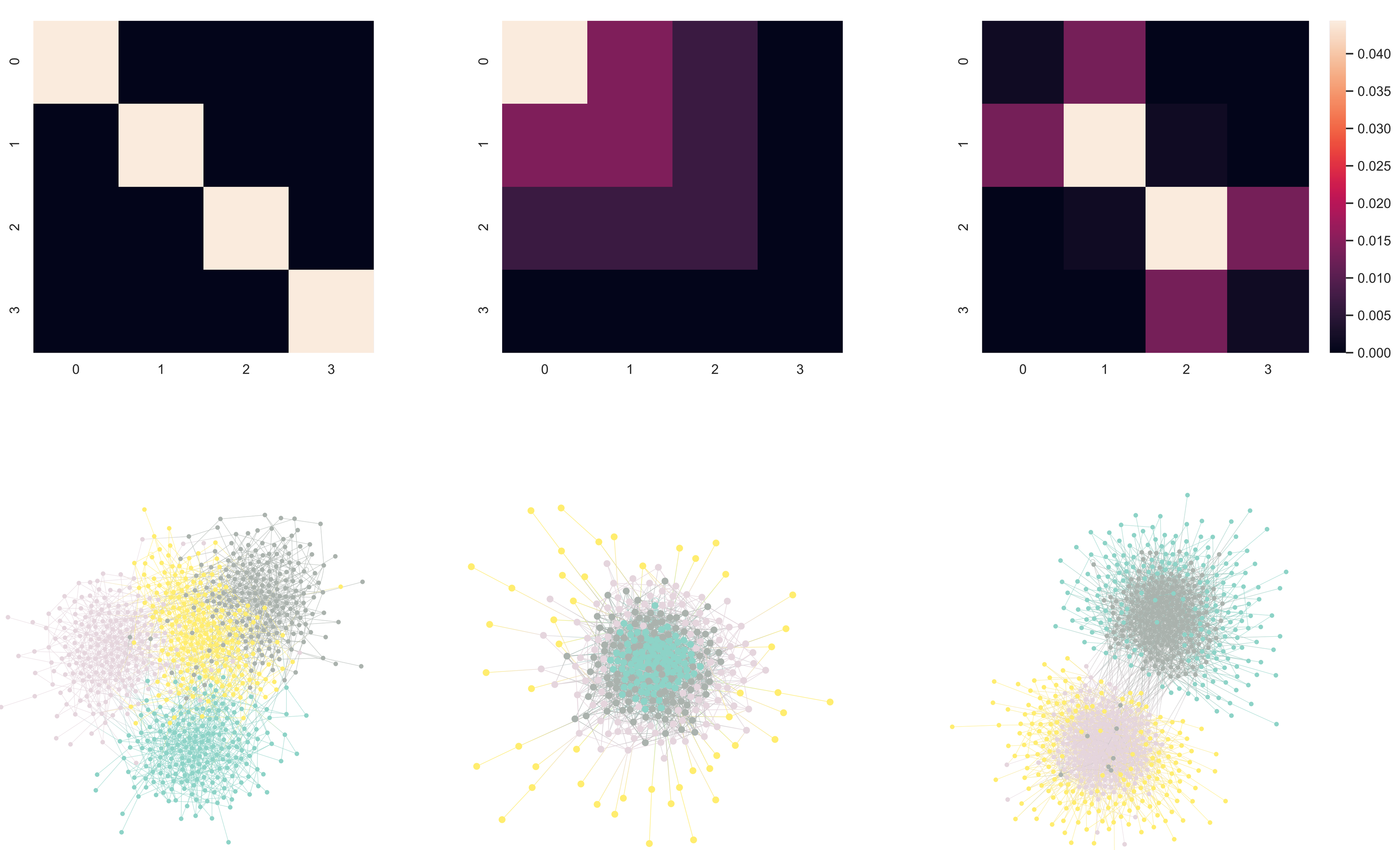
Motivation

- We are interested in finding **assortative** structure in networks

- The general stochastic block models (**SBMs**) can describe general mixing patterns including assortativity as a special case

- However, when assortativity is indeed the dominating pattern, the general model gives more than we need

Our contribution: develop a nonparametric Bayesian approach based on a constrained variant of SBM to detect assortative structure



Top panel: connection matrix indicating the probability of placing edges between different groups with assortative (left) core-periphery (middle) and a mixture of the former two structures (right). Bottom panel: networks generated from SBMs with community structure.

Bayesian inference for community detection

- For an observed network with adjacency matrix A, we sample or maximise from the posterior distribution of vertices partition $\boldsymbol{b}$

- The Bayes' rule

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})}$$

- The marginal likelihood of our model, the planted partition model (**PPM**), reads as

$$P(\boldsymbol{A}|\boldsymbol{b}) = \frac{e_{in}!e_{out}!}{\left(\frac{B}{2}\right)^{e_{in}}\binom{B}{2}^{e_{out}}(E+1)^{1-\delta_{B,1}}} \times \prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \times \frac{\prod_i k_i}{\prod_{i<j} A_{ij}!\prod_i A_{ii}!!}$$

- With any appropriate choice of the prior $P(\boldsymbol{b})$, we can approximate the posterior distribution $P(\boldsymbol{b}|\boldsymbol{A})$ via smapling with Markov Chain Monte Carlo (MCMC)

- For model selection, we compute the description length of the data

$$\Sigma = -\ln P(\boldsymbol{A}|\boldsymbol{b}) - \ln P(\boldsymbol{b})$$

Modularity optimisation and maximum likelihood are not equivalent

- As shown in the literature, there is a connection between the log-likelihood function of PPM and the modularity function

$$\text{Log-likelihood: } \ln\mathcal{L} = \frac{\mu}{2}(A_{ij} - \gamma\theta_i\theta_j)\delta_{b_ib_j} + E\log\lambda_{out} - \frac{\lambda_{out}}{2}\left(\sum_i \theta_i\right)^2 + \sum_i k_i \ln\theta_i$$

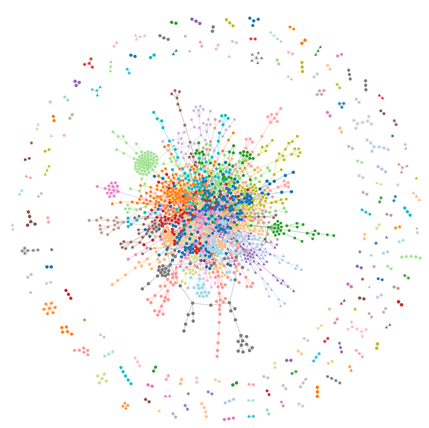
$$\text{Modularity: } Q = \frac{1}{2E}\sum_{ij}\left(A_{ij} - \frac{k_ik_j}{2E}\right)\delta_{b_ib_j}$$

- Maximising modularity is equivalent to maximising the log-likelihood of PPM when the model parameters are set to constant (i.e. $\mu, \gamma, \lambda_{out}$, and $\{\theta_i\}$)

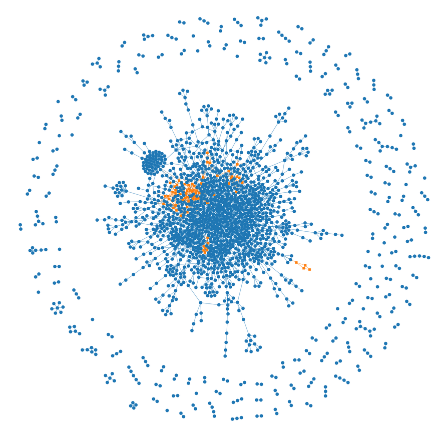
- **However**, such equivalence is tenuous since model parameters should be estimated via the maximum likelihood principle
- Even when it holds, modularity optimisation is prone to overfitting just as the maximum likelihood approach does

Robust against overfitting

Original network

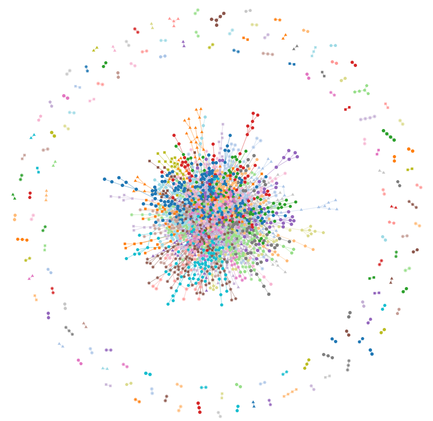


Modularity Optmisation
$B = 185, Q = 0.84$

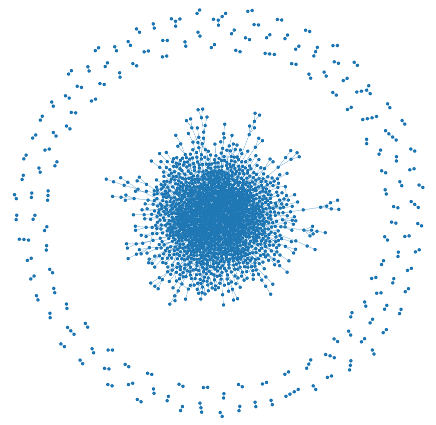


Planted Partition Model
$B = 2, Q = 0.84$

Randomised network



Modularity Optimisation
$B = 168, Q = 0.75$



Planted Partition Model
$B = 2, Q = 0.11$

- We applied modularity optimisation and our Bayesian approach with PPM to a network of protein-protein interactions

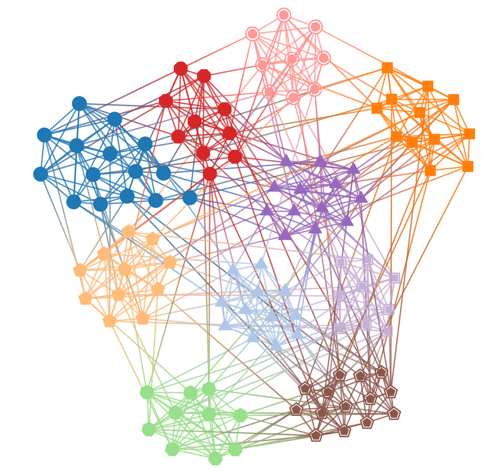
- Results obtained in the original network (top panel) and a randomised version of the network (bottom panel) are shown above

- Modularity optimisation finds over a hundred of communities in the original and the randomised network, **with high value of modularity in both cases**

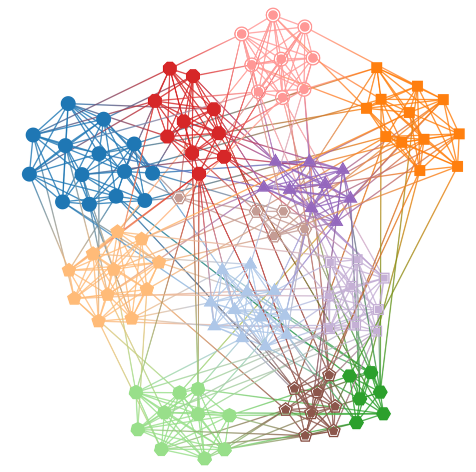
- In comparison, the Bayesian approach **does not** return spurious communities in the random case

Model selection

College football network

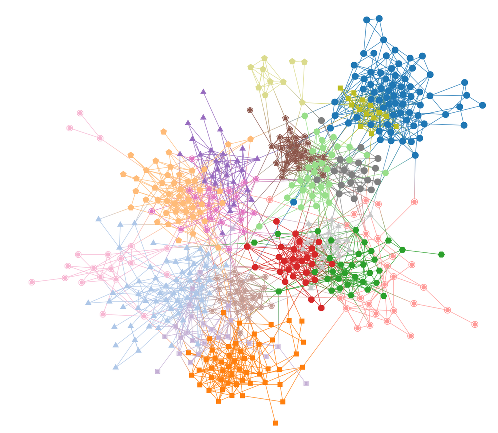


Nested SBM, $\Sigma = 1780.58$

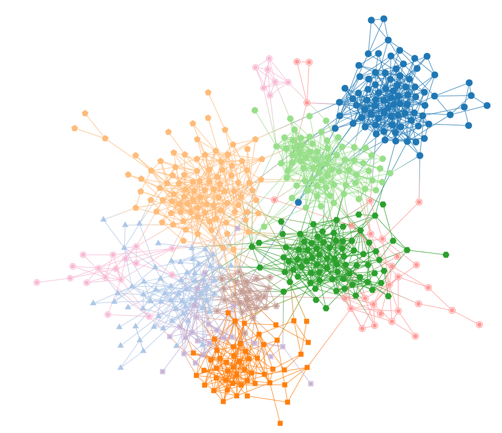


PPM, $\Sigma = 1761.50$

High-school social network



Nested SBM, $\Sigma = 8775.82$

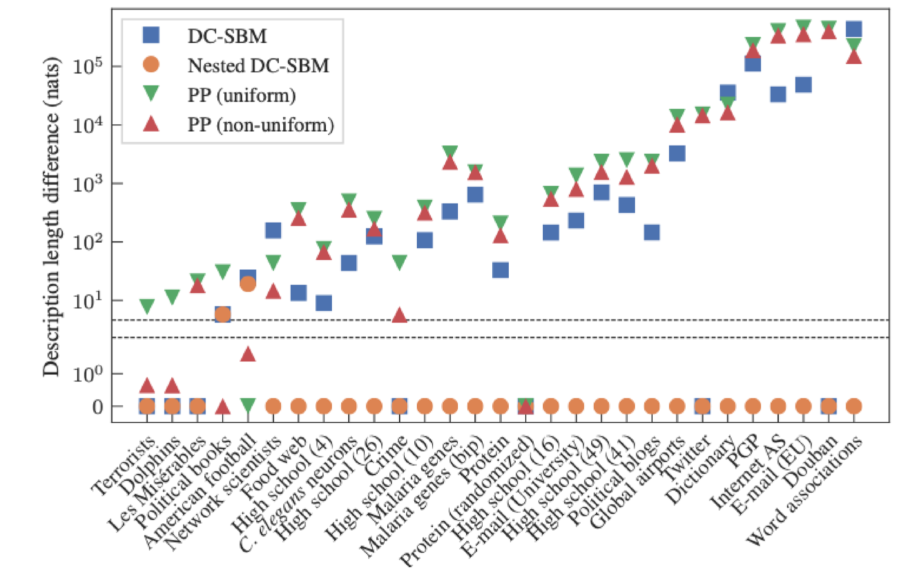


PPM, $\Sigma = 8944.09$

Rather taking for granted, we can check the assumption of assortivity by model selection. The best model is the one has the **minimum description length** Σ

- If assortativity is indeed the dominating pattern, partitions given by PPM should ascribe the smallest Σ compared to other model variants (e.g. college football network)

- When more general pattern is the dominating pattern, other model variants allowing a general mixing pattern should outperform PPM (e.g. high-school social network)



In our study of a selection of empirical networks,

- Only **a few** networks with assortativity being the dominating pattern

- Most of the time (especially in large networks) more general mixing pattern are preferred, raising a **caveat** on the practice of exclusively searching for assortative structure

Further materials

paper: available on arXiv https://arxiv.org/abs/2006.14493
code: available in the *graph-tool* library https://graph-tool.skewed.de/
contact: l.zhang@bath.ac.uk