Inferring community characteristics in labelled networks

IIB Project

Lawrence Tray

Ioannis Kontoyiannis

June 7, 2021

Overview

Introduction

Let's define some terms

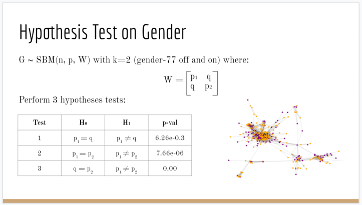
Inferring community characteristics in labelled networks

Network	set of vertices (aka nodes) connected by edges.
Graph	same as a network.
Labelled	information about each vertex. We call these *features*.
Community	Densely connected subset of nodes

Aim to identify what *features* impact graphical structure

How we got here

- Started by performing hypothesis tests on features



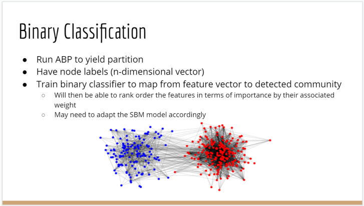


Figure: Slides from Michaelmas presentation

- This proved unhelpful as:
  - Almost all features statistically significant
  - Hard to rank order features against one another
  - Often poor model fit

Instead: find most “natural” partition and see if features can explain

Preliminaries

The stochastic block model (SBM)

We adopt the degree-corrected (DC) microcanonical (MC) formulation [7].

- N – number of vertices
- B – number of blocks

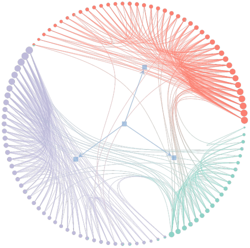


Figure: Typical SBM graph

DC-SBM$_{\mathrm{MC}}$ parameters:

- b – block membership vector
- e – block connectivity matrix
- k – degree sequence

$$A \sim \mathrm{DC\text{-}SBM_{MC}}(b, e, k)$$

With edges placed uniformly at random but respecting:

$$e_{rs} = \sum_{i,j \in [N]} A_{ij} \mathbb{1}\{b_i = r\} \mathbb{1}\{b_j = s\} \qquad \text{and} \qquad k_i = \sum_{j \in [N]} A_{ij}.$$

The feature-first block model

The feature-first block model (FFBM)

Define feature matrix $X \in \{0,1\}^{N \times D}$, this gives us that:

$$X \xrightarrow{\theta} \boxed{b} \xrightarrow{\psi} A$$

Figure: The feature-first block model (FFBM)

Likelihoods

$$p(b|X;\theta) = \prod_{i \in [N]} \phi_{b_i}(x_i;\theta)$$

$$p(A|b;\psi) \sim \mathrm{DC\text{-}SBM}_{\mathrm{MC}}(b, \psi_e, \psi_k)$$

Priors

$$p(\theta) = \mathcal{N}(\theta; 0, \sigma_\theta^2 I)$$

$$p(\psi|b) = p(\psi_e|b)p(\psi_k|\psi_e, b)$$

$$\phi_j(x;\theta) := \frac{\exp(w_j^T x_i)}{\sum_{k \in [B]} \exp(w_k^T x_i)}$$

Important property of the FFBM

Theorem

Our prior choice for $p(\theta)$ gives us that,

$$p(b|X) = B^{-N}.$$

Proof:

$$p(b|X) = \int p(b|X,\theta)p(\theta)d\theta = \int \prod_{i\in[N]} \phi_{b_i}(x_i;\theta)p(\theta)d\theta$$

$$= \prod_{i\in[N]} \int \frac{\exp(w_{b_i}^T x_i) \prod_{j\in[B]} \mathcal{N}(w_j; 0, \sigma_\theta^2 I)}{\sum_{k\in[B]} \exp(w_k^T x_i)} dw_{1:B}.$$

Which is a constant w.r.t. b.

Inference

Inference procedure

We want to draw:

$$\theta^{(t)} \sim p(\theta|A, X).$$

But computing $p(A|\theta, X)$ is $O(B^N) \Rightarrow$ split into:

$$b^{(t)} \sim p\Big(b|A, X\Big)$$
$$\theta^{(t)} \sim p\Big(\theta|X, b^{(t)}\Big)$$

Both steps implemented through Metropolis-Hastings, with:

$$\pi_z(z) - \text{target} \qquad q_z(z, z') - \text{proposal} \qquad \alpha_z(z, z') - \text{accept prob}$$

This is a pseudo-marginal approach. Andrieu and Roberts [1] show that:

$$\theta^{(t)} \sim \mathbb{E}_{b^{(t)}}\left[p(\theta|X, b^{(t)})\right] = p(\theta|A, X)$$

Sampling sequence

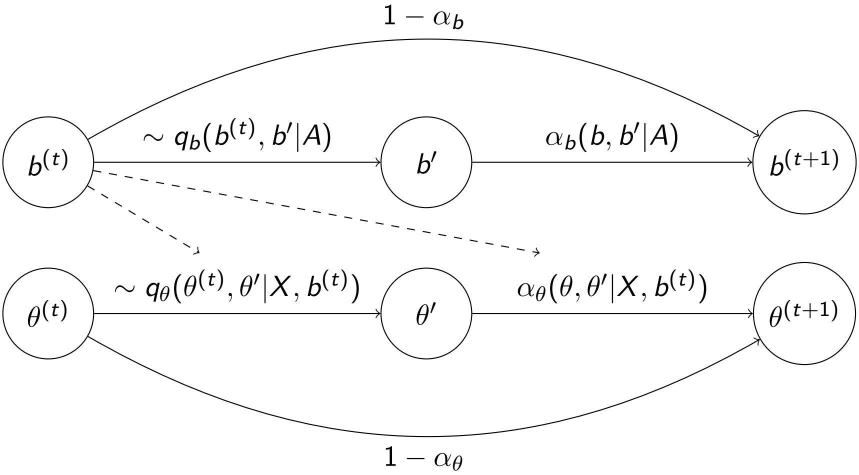
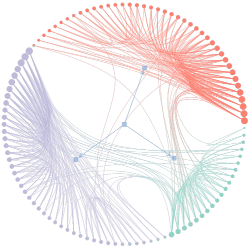


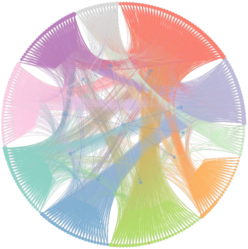
Figure: Sampling sequence.

Experiments

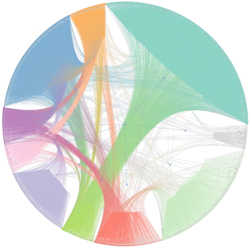
The datasets



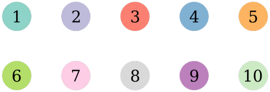
(a) Polbooks $D = 3$



(b) school $D = 13$



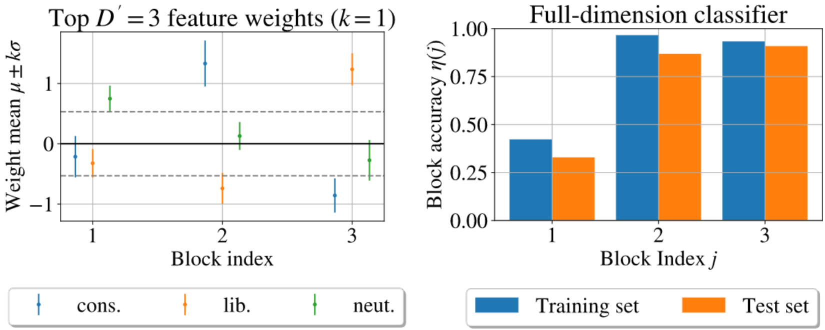
(c) FB egonet $d = 480$



(d) Legend

Figure: Networks laid out and coloured according to inferred block memberships $\hat{y}$ for a given experiment iteration. Visualisation performed using *graph-tool* [6].

Political books [5]



B	D	$\bar{\mathcal{L}}_0$	$\bar{\mathcal{L}}_1$
3	3	0.563 ± 0.042	0.595 ± 0.089

Conclusion

Conclusion

What have we achieved:

- Flip thinking of how we test for a feature's impact on graphical structure
- Developed efficient inference algorithm

What is yet to come:

- FFBM can only explain macro-structure $\Rightarrow$ extend to hierarchical structure

If I could start again

Thanks for listening

Figure: Source, Floyd [2]

References

[1] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697 – 725, 2009. doi: 10.1214/07-AOS574. URL https://doi.org/10.1214/07-AOS574.

[2] The Australian Pink Floyd. Any questions. URL https://www.facebook.com/australianpinkfloyd/posts/if-youve-got-any-questions-youd-like-to-ask-the-band-well-be-doing-a-qa-session-/1643211382381105/.

[3] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444. URL http://www.jstor.org/stable/2334940.

[4] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/7a614fd06c325499f1680b9896beedeb-Paper.pdf.

[5] Boris Pasternak and Ivor Ivask. Four unpublished letters. *Books Abroad*, 44(2):196–200, 1970. ISSN 00067431. URL http://www.jstor.org/stable/40124305.

[6] Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014. doi: 10.6084/m9.figshare.1164194. URL http://figshare.com/articles/graph_tool/1164194.

[7] Tiago P. Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95 (1), Jan 2017. ISSN 2470-0053. doi: 10.1103/physreve.95.012317. URL http://dx.doi.org/10.1103/PhysRevE.95.012317.

[8] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS ONE*, 6(8):1–13, 08 2011. doi: 10.1371/journal.pone.0023176. URL https://doi.org/10.1371/journal.pone.0023176.

Metropolis-Hastings [3]

We want to draw samples $\{z^{(t)}\}$ from some distribution,

$$\pi^*(z) \propto \pi(z).$$

Just need to be able to evaluate $\pi(z)$ point-wise and simulate from a proposal $q(z, z')$. If we accept each proposal with probability,

$$\alpha(z, z') = \min\left(\frac{\pi(z')q(z', z)}{\pi(z)q(z, z')}, 1\right),$$

then the resulting Markov chain is in detailed balance with $\pi(z)$.

b-step

Our target,

$$p(b|A, X) \propto p(b|X)p(A|b) = \pi_b(b),$$

can be evaluated as,

$$\begin{aligned}\pi_b(b) &= p(b|X) \sum_{\psi} p(A, \psi|b) \\ &= p(b|X)p(A, \psi^*|b) \\ &= p(A|\psi^*, b)p(\psi^*|b)p(b|X),\end{aligned}$$

where ψ^* is the only value compatible with (A, b). Big win for the microcanonical formulation.

We borrow $q_b(b, b')$ from Peixoto [7].

θ-step

Our target can be written as:

$$p(\theta|X,b) \propto p(b|X,\theta)p(\theta) = \pi_\theta(\theta) \propto \exp(-U(\theta)).$$

$\therefore$ can write -ve log target as:

$$U(\theta) = \sum_{i,j} y_{ij} \log \frac{1}{a_{ij}} + \frac{1}{2\sigma_\theta^2}||\theta||^2 = N\mathcal{L}(\theta) + \frac{1}{2\sigma_\theta^2}||\theta||^2,$$

where $y_{ij} := \mathbb{1}\{b_i = j\}$ and $a_{ij} = \phi_j(x_i;\theta)$. This form looks very familiar? We can use ∇U to bias our proposal:

$$\theta' = \theta^{(t)} - h_t \nabla U\left(\theta^{(t)}\right) + \sqrt{2h_t} \cdot \xi$$

This is now called the Metropolis-adjusted Langevin algorithm (MALA).

Serialise the chains

The $b^{(t)}$ does not use $\theta^{(t)}$.

The $\theta^{(t)}$ update uses $b^{(t)}$ through $y_{ij}^{(t)} := \mathbb{1}\{b_i^{(t)} = j\}$.

Why not use empirical mean of this quantity:

$$\hat{y}_{ij} := \frac{1}{|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} y_{ij}^{(t)}.$$

This is an unbiased estimate of $p(b_i = j | A, X)$.

Using $\hat{y}$ instead of $y^{(t)}$ for the θ-step:

- Reduced variance in evaluation of U and ∇U
- Can run b and θ-chains sequentially rather than in parallel $\Rightarrow$ different lengths.

Dimensionality Reduction

Want to know which features we can discard:

- Write θ as matrix W, so that W_{ij} is weight for block i and feature j.
- Compute mean $\hat{\mu}_{ij}$ and std dev $\hat{\sigma}_{ij}$ of the $W_{ij}^{(t)}$-samples

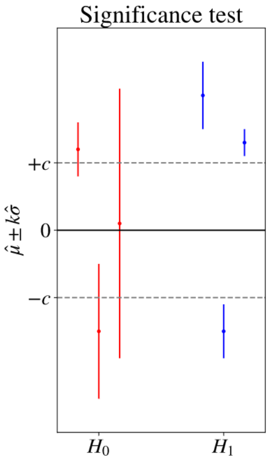
Imagine a test on W_{ij} such that:

$$H_0 : |W_{ij}| \leq c \qquad H_1 : |W_{ij}| > c$$

If we use Laplace approximation can come up with simple decision rule:

$$h_{ij} = H_1 \iff (\hat{\mu}_{ij} - k\hat{\sigma}_{ij}, \hat{\mu}_{ij} + k\hat{\sigma}_{ij}) \cap (-c, +c) = \emptyset$$

With $k > 0$ controlling degree of significance of result



Dimensionality Reduction cont...

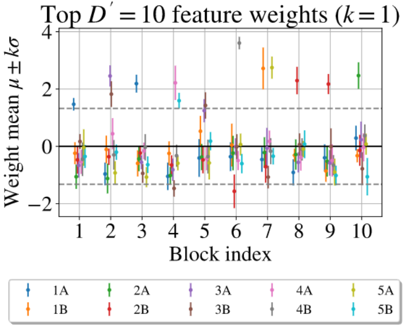
If we specify cut-off $c > 0$ and multiplier $k > 0$ can only retain features d such that:

$$\mathcal{D}' := \{d \in [D] : \exists i \in [B] \text{ s.t. } h_{id} = H_1\}$$

Instead it is often more practical to fix $|\mathcal{D}'| = D'$ and $k = k_0$, then find the maximal cut-off:

$$c^* = \arg\max_{c>0}\{c : |\mathcal{D}'| = D', k = k_0\}$$

Primary school dynamic contacts [8]

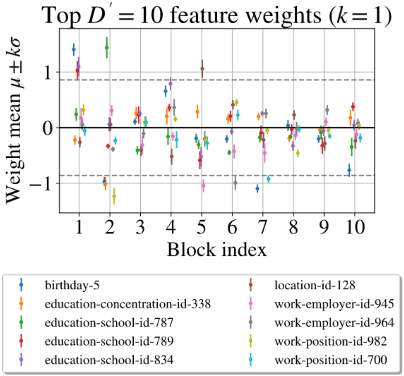


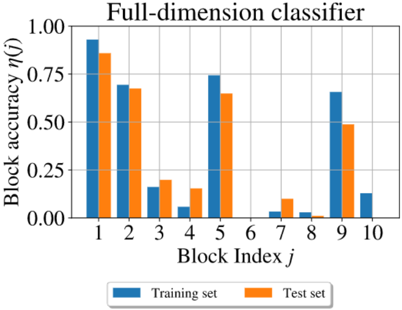
(a) θ-samples. Dotted line is $\pm c^*$.

B	D	D'	$\bar{\mathcal{L}}_0$	$\bar{\mathcal{L}}_1$	$\bar{\mathcal{L}}'_0$	$\bar{\mathcal{L}}'_1$
10	13	10	0.787 ± 0.127	0.885 ± 0.129	0.793 ± 0.132	0.853 ± 0.132

Table: Goodness of fit

Facebook egonet [4]





B	D	D'	$\bar{\mathcal{L}}_0$	$\bar{\mathcal{L}}_1$	$\bar{\mathcal{L}}'_0$	$\bar{\mathcal{L}}'_1$
10	480	10	1.326 ± 0.043	1.538 ± 0.069	1.580 ± 0.150	1.605 ± 0.106

Table: Goodness of fit