The Feature-First Block Model

Lawrence Tray[1], Ioannis Kontoyiannis[2]

[1] Department of Engineering, University of Cambridge, UK
[2] Statistical Laboratory, University of Cambridge, UK

E-mail for correspondence: lpt30@cantab.ac.uk

Abstract: Labelled networks are an important class of data, naturally appearing in applications in science and engineering. A typical inference goal is to determine how the vertex labels (or *features*) affect the network's structure. We introduce a new generative model, the feature-first block model (FFBM), that facilitates the use of rich queries on labelled networks. We develop a Bayesian framework and devise a two-level Markov chain Monte Carlo approach to efficiently sample from the posterior distribution of the FFBM parameters. This allows us to infer if and how the observed vertex-features affect macro-structure. We apply the proposed methods to several real-world networks to extract the most important features along which the vertices are partitioned. Importantly, the whole feature-space is used automatically and features can be rank-ordered implicitly by importance. [The full version of this paper is available on arXiv as [cs.LG] 2105.13762.]

Keywords: Stochastic Block Model; Labelled Networks; Inference.

1 Introduction

Many real-world networks exhibit strong community structure, with most nodes belonging to densely connected clusters. In this work, we examine vertex-labelled networks, referring to the labels as *features*. A typical goal is to determine whether a given feature impacts graphical structure. Answering this requires a random graph model; the standard is the stochastic block model (SBM), see Peixoto (2017).
Analysing a labelled network with one of the standard SBM variants requires partitioning the graph into blocks grouped by distinct values of the feature of interest. The associated model can then be used to test for evidence of heterogeneous connectivity between the feature-grouped blocks. But this approach can only consider disjoint feature sets and the feature-grouped blocks often provide an unnatural partition.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

We would instead prefer to partition the graph into its most natural blocks and then find which of the available features – if any – best predict the resulting partition. Thus motivated, we present a novel framework for modelling labelled networks. This is not the first extension of the SBM to labelled networks, e.g. Stanley et al (2019). However, most of the current approaches are focused on leveraging feature information to partition the graph more reliably in the presence of noise. We seek instead to develop a model well suited for inferring how vertex features impact graphical structure and to report our confidence in those conclusions.

2 Feature-First Block Model

We propose a novel generative model for labelled networks, which we call the feature-first block model (FFBM), illustrated in Figure 1. Let N denote the number of vertices, B the number of blocks and D the number of features associated with each vertex. We write X for the $N \times D$ *feature matrix* containing the feature vectors $\{x_i\}_{i=1}^N$ as its rows. For the FFBM, we start with the feature matrix X and generate a random vector of block memberships $b \in [B]^N$, where we write $[K] = \{1, 2 \dots K\}$. For each vertex i, the block membership $b_i \in [B]$ is generated based on the feature vector x_i, independently between vertices, so $p(b|X, \theta) = \prod_{i \in [N]} \phi_{b_i}(x_i; \theta)$.

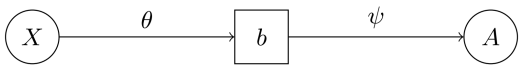

FIGURE 1. The Feature-First Block Model (FFBM)

Once the block memberships b have been generated, we then draw the adjacency matrix of the graph $A \sim \text{DC-SBM}_{\text{MC}}(b, \psi)$ from the microcanonical DC-SBM, Peixoto (2017), with additional parameters ψ. Appropriate priors are placed on the parameters θ and ψ to complete the Bayesian framework.

3 Inference

Given a labelled network (A, X), we wish to infer if and how the observed features X impact the graphical structure A. Formally, this means characterising the posterior distribution for θ, $p(\theta|A, X) \propto p(\theta) \cdot p(A|X, \theta)$. Following standard Bayesian practice, we propose an iterative Markov chain Monte Carlo (MCMC) approach to obtain samples $\theta^{(t)}$ from this posterior. First we sample $b^{(t)}$ from the block membership posterior, and then use $b^{(t)}$ to obtain a corresponding sample $\theta^{(t)}$,

$$b^{(t)} \sim p\big(b|A, X\big) \quad \text{then} \quad \theta^{(t)} \sim p\big(\theta|X, b^{(t)}\big). \tag{1}$$

Splitting the Markov chain into two levels side-steps the intractable summation over all latent $b \in [B]^N$ required to directly compute the likelihood, $p(A|X,\theta)$. The resulting $\theta^{(t)}$ samples are asymptotically unbiased in that the expectation of their distribution converges to the true posterior.

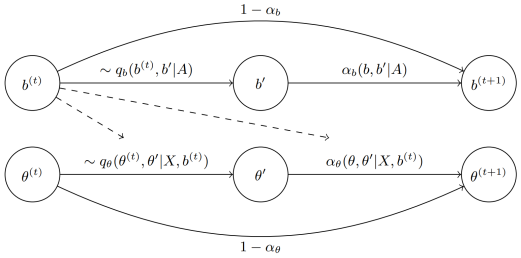

FIGURE 2. θ-sample generation.

Figure 2 shows an overview of the proposed method, with q and α denoting the Metropolis-Hastings proposal distribution and acceptance probability. Due to the formulation of the FFBM, evaluating $p(b|X)$ does not depend on X so we do not need X to sample b. And on the other level, in order to obtain samples for θ we use only b but not A, as $(\theta \perp\!\!\!\perp A)|b$.

3.1 Sampling block memberships

To generate the required b-samples, we adopt the MCMC procedure of Peixoto (2017), which relies on writing the posterior in the following form,

$$p(b|A,X) \propto p(A|b,X) \cdot p(b|X) = \pi_b(b), \tag{2}$$

where $\pi_b(\cdot)$ denotes the un-normalised target density. As we are using the microcanonical SBM formulation, there is only one value of ψ - denoted ψ^* - that is compatible with the given (A,b) pair; see Peixoto (2017). Therefore, the summation needed to evaluate $p(A|b,X)$ reduces to just one term: $p(A|b,X) = \sum_\psi p(A,\psi|b,X) = p(A,\psi^*|b,X)$. In this context, the microcanonical entropy of the configuration b is, $S(b) \triangleq -\log \pi_b(b)$, which can be thought of as the optimal "description length" of the graph; this will later be employed to evaluate experimental results.

3.2 Sampling feature-to-block generator parameters

The target distribution for the required θ-samples is the posterior of θ given the values of the pair (X,b). We write this as,

$$\pi_\theta(\theta) \propto p(\theta|X,b) \propto p(b|X,\theta)p(\theta) \propto \exp\left(-U(\theta)\right), \tag{3}$$

where $U(\theta)$ denotes the negative log-posterior. Let $y_{ij} \triangleq \mathbf{1}\{b_i = j\}$ and $a_{ij} \triangleq \phi_j(x_i; \theta)$. Discarding constant terms, $U(\theta)$ can be expressed as,

$$U(\theta) = \left(\sum_{i \in [N]} \sum_{j \in [B]} y_{ij} \log \frac{1}{a_{ij}} \right) + \frac{1}{2\sigma_\theta^2} \|\theta\|^2 = N \cdot \mathcal{L}(\theta) + \frac{1}{2\sigma_\theta^2} \|\theta\|^2. \quad (4)$$

The function $U(\theta)$ is a typical objective function for neural network training. The first term $N \cdot \mathcal{L}(\theta)$ comes from the likelihood and is the cross-entropy between the graph-predicted and feature-predicted block memberships. The second term, introduced by the prior, brings a form of regularisation, guarding against over-fitting. To draw samples from the posterior $\pi_\theta \propto \exp(-U)$ we adopt the Metropolis-adjusted Langevin algorithm, MALA - Roberts and Tweedie (1996), which uses ∇U to bias the proposal towards regions of higher density.

3.3 Sampling sequence

So far, each $\theta^{(t)}$ update has used its corresponding $b^{(t)}$ sample. This means the evaluation of $U^{(t)}$ and $\nabla U^{(t)}$ has high variance, leading to longer burn-in and possibly slower MCMC convergence. The only link between $b^{(t)}$ and $\theta^{(t)}$ is in the evaluation of $U^{(t)}$ and $\nabla U^{(t)}$ which depends only on the matrix $y^{(t)}$ with entries $y_{ij}^{(t)} \triangleq \mathbf{1}\{b_i^{(t)} = j\}$. We would rather deal with the expectation of each $y_{ij}^{(t)}$: $\mathbb{E}\left[y_{ij}^{(t)}\right] = \mathbb{E}_{b^{(t)}}\left[\mathbf{1}\left(b_i^{(t)} = j\right)\right] = p(b_i = j | A, X)$. An unbiased estimate for this can be obtained using the thinned b-samples after burn-in. Let $\mathcal{T}_b$ denote the retained set of indices for the b-samples and $\mathcal{T}_\theta$ similarly for the θ-chain. The unbiased estimate for $y_{ij}^{(t)}$ is then:

$$\hat{y}_{ij} \triangleq \frac{1}{|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} y_{ij}^{(t)} = \frac{1}{|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} \mathbf{1}\{b_i^{(t)} = j\}. \quad (5)$$

The same matrix $\hat{y}$ is used in each $\theta^{(t)}$ update step. This way, it is not necessary to run the b and θ Markov chains concurrently. Instead, we run the b-chain to completion and use it to generate $\hat{y}$ also allowing us to vary the lengths of each.

3.4 Dimensionality reduction

Evaluating U and ∇U is linear in the dimension of the feature space D, so there is computational incentive to reduce D. Given the samples $\left\{\theta^{(t)}\right\}$, we compute the empirical mean and standard deviation of each component. Switching to the matrix notation W for θ, let:

$$\hat{\mu}_{ij} \triangleq \frac{1}{|\mathcal{T}_\theta|} \sum_{t \in \mathcal{T}_\theta} W_{ij}^{(t)} \qquad \text{and} \qquad \hat{\sigma}_{ij}^2 \triangleq \frac{1}{|\mathcal{T}_\theta|} \sum_{t \in \mathcal{T}_\theta} \left(W_{ij}^{(t)} - \hat{\mu}_{ij}\right)^2. \quad (6)$$

A simple heuristic to discard the least important features requires specifying a cutoff $c > 0$ and a multiplier $k > 0$. We define the function $\mathcal{F}_i(j)$ as in (7) and only keep features with indices $d \in \mathcal{D}'$, where $\mathcal{D}'$ is given in (8).

$$\mathcal{F}_i(j) \triangleq (\hat{\mu}_{ij} - k\hat{\sigma}_{ij}, \hat{\mu}_{ij} + k\hat{\sigma}_{ij}) \cap (-c, +c), \tag{7}$$

$$\mathcal{D}' \triangleq \{j \in [D] : \exists i \in [B] \text{ s.t. } \mathcal{F}_i(j) = \emptyset\}. \tag{8}$$

Intuitively, this means discarding any feature j for which $(\hat{\mu}_{ij} - k\hat{\sigma}_{ij}, \hat{\mu}_{ij} + k\hat{\sigma}_{ij})$ overlaps with $(-c, c)$ for all i. Conversely, if we want to fix the number of dimensions in our reduced feature set $|\mathcal{D}'| = D'$, the problem then becomes finding the largest value of c such that $|\mathcal{D}'| = D'$ given $k = k_0$:

$$c^* = \text{argmax}\{c > 0 \ : \ |\mathcal{D}'| = D', k = k_0\}. \tag{9}$$

4 Experimental results

We apply our proposed methods to a variety of real-world datasets. The inferred partitions b for all of these are given on Figure 3. To assess model performance, the average description length per entity (nodes and edges) $\bar{S}_e$ is used to gauge the SBM fit, and the vertex set $[N]$ is partitioned at random into training and test sets, $\mathcal{G}_0$ and $\mathcal{G}_1$, to assess the performance of the feature-to-block predictor. The average cross-entropy loss over each set, denoted $\bar{\mathcal{L}}_\star$, is used to gauge the quality of the fit.
For higher-dimensional datasets, we develop a novel dimensionality reduction method to select only the top D' features. We then retrain the feature-block predictor using only the retained feature set, and report the cross-entropy loss $\bar{\mathcal{L}}'_\star$ over the training and test sets for the reduced classifier.
Table 1 summarises the results for each experiment. We see that the dimensionality reduction procedure brings the training and test losses closer, indicating that the retained features are indeed well correlated with the underlying graphical partition and that the approach generalises correctly.

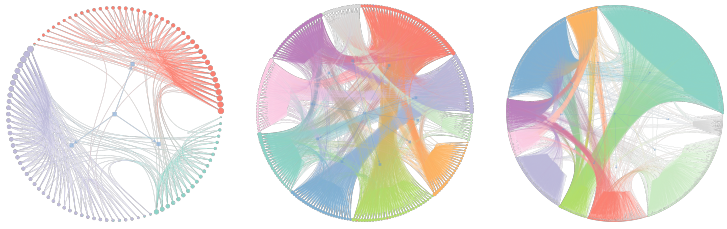

FIGURE 3. Networks laid out and coloured according to inferred block memberships. Left to right: Polbooks, Krebs (2004); Primary School, Stehle et al (2011); Facebook Egonet, Leskovec and Mcauley (2012).

TABLE 1. Results averaged over $n = 10$ iterations (mean $\pm$ std. dev.).

Dataset	B	D	D'	$\bar{S}_e$	$\bar{\mathcal{L}}_0$	$\bar{\mathcal{L}}_1$	c^*	$\bar{\mathcal{L}}'_0$	$\bar{\mathcal{L}}'_1$
Polbooks	3	3	–	2.250 ± 0.000	0.563 ± 0.042	0.595 ± 0.089	–	–	–
School	10	13	10	1.894 ± 0.004	0.787 ± 0.127	0.885 ± 0.129	1.198 ± 0.249	0.793 ± 0.132	0.853 ± 0.132
FB egonet	10	480	10	1.626 ± 0.003	1.326 ± 0.043	1.538 ± 0.069	0.94 ± 0.019	1.580 ± 0.150	1.605 ± 0.106

5 Conclusion

The feature-first block model (FFBM) is introduced, as a new generative model for labelled networks with communities. An efficient MCMC algorithm is developed for sampling from the posterior distribution of the relevant parameters in the FFBM; the main idea is to divide up the graph into its most natural partition under the associated parameter values, and then to determine whether the vertex features can accurately explain the partition. Through applications on empirical network data, this approach is demonstrated to be effective at extracting and describing the most natural communities in a labelled network. Nevertheless, it can only currently explain the structure at the macroscopic scale. Future work will benefit from extending the FFBM to a further hierarchical model, so that the structure of the network can be explained at all scales of interest.

References

Krebs, V. (2004) . The political books network, http://www.orgnet.com/.

Leskovec, J., Mcauley, J. (2012). Learning to discover social circles in ego networks. *Advances in Neural Information Processing Systems* vol. 25

Peixoto, T.P. (2017). Nonparametric Bayesian inference of the microcanonical stochastic block model. Physical Review E 95(1).

Roberts, G.O., Tweedie, R.L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 2(4).

Stanley, N. et al (2019). Stochastic block models with multiple continuous attributes. Applied Network Science, 4(1), 1-22.

Stehle, J. et al (2011). High resolution measurements of face-to-face contact patterns in a primary school. PLoS ONE 6(8), 1—13.

Tray, L., Kontoyiannis, I. (2021). The feature-first block model. arXiv preprint 2105.13762 [cs.LG].