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 numerous applications in science and engineering. A typical inference goal is to determine how the vertex labels (or *features*) affect the network's structure. In this work, 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 relevant 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 a variety of network data to extract the most important features along which the vertices are partitioned. The main advantages of the proposed approach are that the whole feature-space is used automatically and that features can be rank-ordered implicitly according to impact.

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. Finding ways to recover the latent communities from the observed graph is an important task in many applications, including compression [**?**] and link prediction [**?**]. 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) [**?**]. Numerous variants of the SBM have been proposed, e.g., the MMSBM [**?**] and OSBM [**?**], but these do not include features in the graph generation process.

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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

To analyse a labelled network using one of the simple SBM variants, a typical procedure would be to partition 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. Nevertheless, this approach can only consider disjoint feature sets and the feature-grouped blocks are often an unnatural partition of the graph.
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, which we call the feature-first block model (FFBM). This is an extension of the SBM to labelled networks.

2 Feature-First Block Model

In this section we propose a novel generative model for labelled networks. We call this the feature-first block model (FFBM), illustrated in Figure 1. Let N denote the number of vertices, B the number of blocks and $\mathcal{X}$ the set of values each feature can take. We define the vector $x_i \in \mathcal{X}^D$ as the feature vector for vertex i, where D is the number of features associated with each vertex. For example, in the datasets we analyse, we deal with binary feature flags (denoting the presence/absence of each feature), so $\mathcal{X} = \{0, 1\}$. 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 $b \in [B]^N$. For each vertex i, the block membership $b_i \in [B]$ is generated based on the feature vector x_i, independently between vertices. The conditional distribution of b_i given x_i also depends on a collection of weight vectors $\theta = \{w_k\}_{k=1}^B$, where each w_k has dimension D. We will later find it convenient to write θ as a $B \times D$ matrix of weights W. Specifically, the distribution of b given X and θ is,

$$p(b|X, \theta) = \prod_{i \in [N]} p(b_i|x_i, \theta) = \prod_{i \in [N]} \phi_{b_i}(x_i; \theta) = \prod_{i \in [N]} \frac{\exp\left(w_{b_i}^T x_i\right)}{\sum_{k \in [B]} \exp\left(w_k^T x_i\right)}. \tag{1}$$

Note that ϕ_{b_i} has the form of a softmax activation function. More complex models based on different choices for the distributions ϕ_{b_i} above are also possible, but then deriving meaning from the inferred parameter distributions is more difficult.
Once the block memberships b have been generated, we then draw the graph A from the microcanonical DC-SBM with additional parameters $\psi = \{\psi_e, \psi_k\}$:

$$A \sim \text{DC-SBM}_{\text{MC}}(b, \psi_e, \psi_k). \tag{2}$$

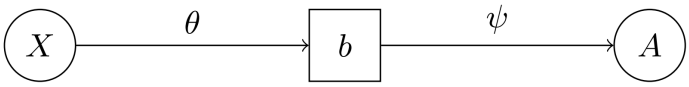

FIGURE 1. The Feature-First Block Model (FFBM)

3 Inference

Having completed the definition of the FFBM, we wish to leverage it to perform inference. Specifically, 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: $p(\theta|A, X) \propto p(\theta)\cdot p(A|X, \theta)$. Although the prior is easily computable, computing the likelihood requires summing over all latent block-states, $p(A|X, \theta) = \sum_{b\in[B]^N} p(A|b)P(b|X, \theta)$, which is clearly impractical. In fact, this approach is doubly intractable as we would also need to compute the normalising constant $p(A|X)$. Therefore, following standard Bayesian practice, instead we aim to draw samples from the posterior,

$$\theta^{(t)} \sim p(\theta|A, X). \tag{3}$$

We propose an iterative Markov chain Monte Carlo (MCMC) approach to obtain these samples $\{\theta^{(t)}\}$. We first draw a sample $b^{(t)}$ from the block membership posterior, and then use $b^{(t)}$ to obtain a corresponding sample $\theta^{(t)}$:

$$b^{(t)} \overset{\text{distr}}{\approx} p\big(b|A, X\big) \quad \text{then} \quad \theta^{(t)} \overset{\text{distr}}{\approx} p\big(\theta|X, b^{(t)}\big), \tag{4}$$

where these approximations become exact as the number of MCMC iterations $t \to \infty$. As described in the following subsections, this can be implemented through a two-level Markov chain via the Metropolis-Hastings (MH) algorithm [**?**]. The splitting of the Markov chain into two levels allows us to side-step the 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:

$$\lim_{t\to\infty} \mathbb{E}_{b^{(t)}} \left[p\left(\theta|X, b^{(t)}\right)\right] = \sum_{b\in[B]^N} p(\theta|X, b)p(b|A, X) = p(\theta|A, X). \tag{5}$$

This is an example of a pseudo-marginal approach; see, e.g., Andrieu and Roberts [**?**] for a detailed rigorous derivation based on (5).
Figure 2 shows an overview of the proposed method, with q and α denoting the MH proposal distribution and acceptance probability respectively. Note the importance of the simplification in (**??**). As evaluating $p(b|X)$ does not depend on X, 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 $(\theta \perp\!\!\!\perp A)|b$.

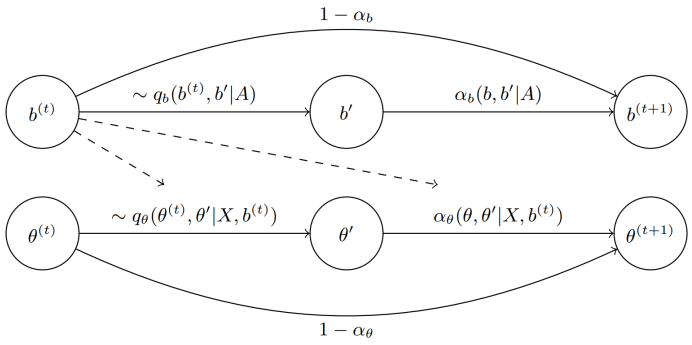

FIGURE 2. θ-sample generation.

4 Experimental results

We apply our proposed methods to a variety of labelled networks:

- **Political books** [**?**] ($N = 105, E = 441, D = 3$) – network of Amazon political books published close to the 2004 presidential election. Two books are connected if they were frequently co-purchased. Vertex features encode the political affiliation of the author (liberal, conservative, or neutral).
- **Primary school dynamic contacts** [**?**] ($N = 238, E = 5539, D = 13$) – network of 238 individuals (students and teachers), with edges denoting face-to-face contacts at a primary school in Lyon, France. The vertex features are class membership (one of 10 values: 1A-5B), gender (male, female), and status (teacher, student). We choose to analyse just the second day of results.
- **Facebook egonet** [**?**] ($N = 747, E = 30025, D = 480$) – network of Facebook users with edges denoting "friends". Vertex features are fully anonymised and encode information about each user's education history, languages spoken, gender, home-town, birthday etc. We focus on the egonet with id 1912.

For reference, the inferred partitions for all of these are given on Figure 3. We employ the following metrics to assess model performance. First, the average description length per entity (nodes and edges) $\bar{S}_e$ used to gauge the SBM fit is defined as:

$$\bar{S}_e \triangleq \frac{1}{(N+E)|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} S\left(b^{(t)}\right). \tag{6}$$

Next, to assess the performance of the feature-to-block predictor, the vertex set $[N]$ is partitioned at random so that a constant fraction f of vertices

form the training set $\mathcal{G}_0$ and the remainder form the test set $\mathcal{G}_1$. The b-chain is run using the whole network but only vertices $v \in \mathcal{G}_0$ are used for the θ-chain. Then the the average cross-entropy loss over each set is used to gauge the quality of the fit,

$$\bar{\mathcal{L}}_\star \triangleq \frac{1}{|\mathcal{T}_\theta|} \sum_{t \in \mathcal{T}_\theta} \mathcal{L}_\star^{(t)}, \quad \text{where} \quad \mathcal{L}_\star^{(t)} \triangleq \frac{1}{|\mathcal{G}_\star|} \sum_{i \in \mathcal{G}_\star} \sum_{j \in [B]} \hat{y}_{ij} \log \frac{1}{\phi_j\left(x_i; \theta^{(t)}\right)}, \tag{7}$$

where $\star \in \{0, 1\}$ toggles between the training and test sets and $\hat{y}_{ij}$ is defined in (**??**). Nevertheless, the cross-entropy loss is a coarse measure of fit. A new measure, specific to each detected block, can be defined as follows. Let $\mathcal{B}_\star(j)$ be the set of vertices with maximum a posteriori probability of belonging to block j, $\mathcal{B}_\star(j) \triangleq \{i \in \mathcal{G}_\star : \hat{b}_i = j\}$, where $\hat{b}_i \triangleq \underset{j}{\text{argmax}} \hat{y}_{ij}$, and define the *block-accuracy* for block j as,

$$\eta_\star(j) \triangleq \frac{1}{|\mathcal{B}_\star(j)| \cdot |\mathcal{T}_\theta|} \sum_{i \in \mathcal{B}_\star(j)} \sum_{t \in \mathcal{T}_\theta} \mathbf{1}\left\{\hat{b}_i = \underset{j}{\text{argmax}} \phi_j\left(x_i; \theta^{(t)}\right)\right\}. \tag{8}$$

This effectively tests whether the feature-to-block and graph-to-block predictions agree in their largest component. For the higher-dimensional datasets, we also apply the dimensionality reduction method of Section **??**. We then retrain the feature-block predictor using only the retained feature set $\mathcal{D}'$, and report the log-loss over the training and test sets for the reduced classifier – denoted $\bar{\mathcal{L}}'_0$ and $\bar{\mathcal{L}}'_1$ respectively.

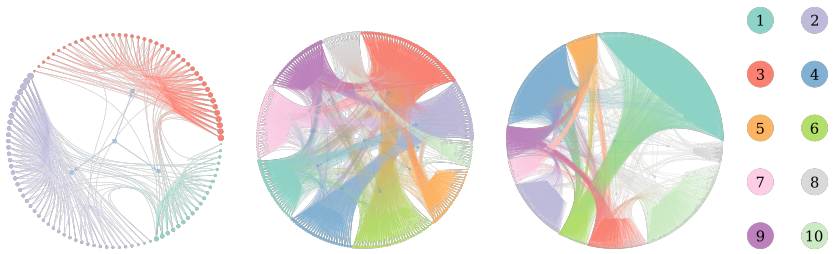

FIGURE 3. Networks laid out and coloured according to inferred block memberships for a given experiment iteration. Visualisation performed using `graph-tool` [**?**].

TABLE 1. Experimental 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 proposed Feature-First Block Model (FFBM) is a new generative model for labelled networks. It is a hierarchical Bayesian model, well-suited for describing how features affect network structure. The Bayesian inference tools developed in this work facilitate the identification of vertex features that are in some way correlated with the network's graphical structure. Consequently, finding the features that best describe the most pronounced partition, makes it possible in practice to examine the existence of – and to make a case for – causal relationships.
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 several applications on empirical network data, this approach is shown 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

Abbe, E. (2016) Graph compression: The effect of clusters. In: *54th Annual Allerton Conference on Communication, Control, and Computing.* pp. 1—8.

Airoldi, E.M., Blei, D., Fienberg, S., Xing, E. (2009). Mixed membership stochastic blockmodels. In: *Advances in Neural Information Processing Systems* vol. 21.

Andrieu, C., Roberts, G.O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics 37(2), 697—725.

Gaucher, S., Klopp, O., Robin, G. (2021). Outliers detection in networks with missing links. Computational Statistics & Data Analysis 164, 107308.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1), 97-–109.

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

Nowicki, K., Snijders, T.A.B. (2001). Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96(455), 1077—1087.

Pasternak, B., Ivask, I. (1970). Four unpublished letters. Books Abroad 44(2), 196—200.

Peixoto, T.P. (2014). Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. Physical Review E 89(1).

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

Peixoto, T. (2014). The graph-tool Python library. figshare (2014), figshare.com/articles/graph_tool/1164194

Roberts, G.O., Tweedie, R.L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 2(4), 341—363.

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

Zhu, J., Song, J., Chen, B. (2016). Max-margin nonparametric latent feature models for link prediction. arXiv preprint cs.LG:1602.07428.