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. 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) [Nowicki and Snijders (2001)].
INSERT REBUTTAL OF CURRENT METHODS
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

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.

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

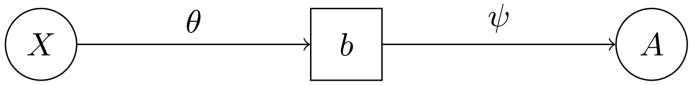

FIGURE 1. The Feature-First Block Model (FFBM)

Once the block memberships b have been generated, we then draw the adjacency matrix of the graph, A, from the microcanonical DC-SBM [Peixoto (2017)] with additional parameters ψ,

$$A \sim \text{DC-SBM}_{\text{MC}}(b, \psi). \tag{1}$$

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: $p(\theta|A, X) \propto p(\theta) \cdot p(A|X, \theta)$. Therefore, following standard Bayesian practice, instead we aim to draw samples from the posterior,

$$\theta^{(t)} \sim p(\theta|A, X). \tag{2}$$

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 $\theta^{(t)}$:

$$b^{(t)} \overset{\text{distr}}{\approx} p\big(b|A,X\big) \quad \text{then} \quad \theta^{(t)} \overset{\text{distr}}{\approx} p\big(\theta|X,b^{(t)}\big), \tag{3}$$

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.

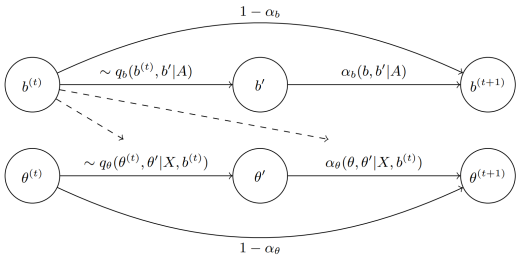

FIGURE 2. θ-sample generation.

Figure 2 shows an overview of the proposed method, with q and α denoting the Metropolis-Hastings proposal distribution and acceptance probability respectively. Due to the FFBM's formulation, 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 $(\theta \perp\!\!\!\perp A)|b$.

4 Experimental results

We apply our proposed methods to a variety of real-world datasets. For reference, the inferred partitions for all of these are given on Figure 3.

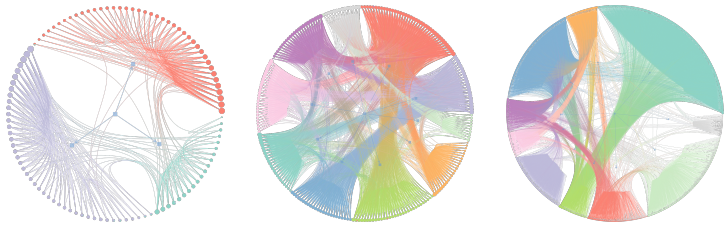

FIGURE 3. Networks laid out and coloured according to inferred block memberships. Left to right: Polbooks, Krebs (2004); Primary School, Stehle et al (2011); Facebook Egonet, Leskovec and Mcauley (2012).

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

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

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

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

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