UNIVERSITY OF CAMBRIDGE
Department of Engineering

Inferring community characteristics in labelled networks

Author Name: Lawrence Tray

Supervisor: Ioannis Kontoyiannis

Date: 02/06/21

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

Signed: *Lawrence Tray* *date:* 02/06/21

Technical Abstract

Labelled networks form a very common and 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.

A standard approach has been to partition the network into blocks grouped by distinct values of the feature of interest. A block-based random graph model – typically a variant of the stochastic block model – is then used to test for evidence of asymmetric behaviour between these feature-based communities. Nevertheless, the resulting communities often do not produce a natural partition of the graph.

In this work, we introduce a new generative model, the feature-first block model (FFBM), which is more effective at describing vertex-labelled undirected graphs and also facilitates the use of richer queries on labelled networks. We develop a Bayesian framework for inference with this model, and we present a method to efficiently sample from the posterior distribution of the FFBM parameters. The FFBM's structure is kept deliberately simple to retain easy interpretability of the parameter values.

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. Any features that do not significantly impact the high-level structure can be discarded to reduce the problem dimension. In cases where the vertex features available do not readily explain the community structure in the resulting network, the approach detects this and is protected against over-fitting.

Future work on this topic could benefit from extending the FFBM to be hierarchical in nature. Such a modification would be relatively natural and would increase the descriptive power of the model. The current FFBM formulation is limited in that it can only explain structure at the macro-scale. Nevertheless, a hierarchical formulation would allow explanations of structure at any arbitrary length-scale.