Inferring community characteristics in labelled networks

Lawrence Tray
Department of Engineering
University of Cambridge
lpt30@cam.ac.uk

Ioannis Kontoyiannis
Department of Mathematics
University of Cambridge
ik355@cam.ac.uk

Abstract

Labelled networks are an extremely common and important form of data. A typical inference goal is to determine how the vertex labels (called features) affect graphical structure. The standard approach to this problem 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 (SBM) - is then used to test for evidence of asymmetric behaviour within these feature-based communities.

Nevertheless, these feature-based communities are often not a natural partition of the graph and thus the models employed are rarely a good fit. With this in mind, we present a novel generative model, which we call the feature-first block model (FFBM), for better describing vertex-labelled undirected graphs. This allows us to perform richer queries on labelled networks. We present a method to efficiently sample the FFBM parameters for inference. The FFBM's structure is kept deliberately simple to retain easy interpretability of the parameter values.

We apply the developed methods to a variety of network data to extract the most important features along which the vertices divide themselves. The greatest advantage of the proposed approach is that the whole feature-space is used automatically and features can be rank-ordered implicitly according to impact. Any features that do not greatly impact the high-level structure can be discarded to reduce the problem dimension. In the case the vertex features available do not readily explain the community structure in the resulting network, the approach detects this and is protected against over-confidence. Future work may benefit from extending the FFBM to multiple hierarchical levels. This would allow the structure to be explained at each level of coarseness rather than simply at the highest level.

1 Introduction

A somewhat surprising property of many real-world networks is that they exhibit strong community structure. In other words, each node will often belong to a cluster of densely connected nodes. There is high interest in recovering the latent communities from the observed graphs. The inferred communities can be exploited for compression algorithms [1] or used for link prediction in incomplete networks [5] to name but a few applications.

We restrict our analysis to labelled networks. These are graphs where we additionally have information about the properties of each vertex. We shall refer to these vertex properties as features. One of the most common questions we can ask of labelled networks is whether a given vertex feature has an impact on the structure of the graph. To answer this question from a Bayesian perspective we must use a random graph model; the most common form is called the stochastic block model (SBM) [10]. This is a latent variable model where each vertex belongs to a single block and the probability two

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

nodes are connected depends only on the block memberships of each. There have been many variants to this model; the most popular are the mixed-membership stochastic block model (MMSBM) [2] and the overlapping stochastic block model (OSBM) [19]. Effectively, these just extend the model to allow each vertex to belong to multiple blocks simultaneously.

However, a major drawback of these graphical models as applied to labelled networks is that they do not automatically include vertex features in the random graph generation process. Approaches based on graph neural networks [9] that utilise vertex features have been developed but these lack the easy interpretability of the simpler models.

To analyse a labelled network using one of the simple SBM variants, a typical inference 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 is limited in that it can only consider one feature at a time; this means any conclusions drawn highly vulnerable to the presence of confounding variables. Furthermore, considering each feature separately makes it difficult to rank order the features by magnitude of impact. Lastly, the feature-grouped blocks are often an unnatural partition of the graph, leading to a poor model fit. We would instead prefer to partition the graph into its most natural blocks and then determine which features best predict the resulting partition.

With these desiderata in mind, we present a novel framework for modelling labelled networks, which we call the feature-first block model (FFBM). This can be thought of as an extension of the SBM to labelled networks. In the FFBM, we use the features first to generate the latent block membership for each vertex. The latent block membership is therefore a stepping stone rather than a starting point in our analysis. We go on to present an efficient algorithm for sampling from the parameters of the feature-to-block generator. We can interpret the sampled FFBM parameters to determine which features have the largest impact on overall graphical structure. Any features found to be irrelevant can be discarded as a form of dimensionality reduction.