Community Identification in Weighted Networks

Dan Kessler

April 29, 2019

Abstract

In this report we explore the problem of community identification in weighted networks under the stochastic block model (SBM). First, we introduce some important notation and ideas regarding graphs and networks and then providing background on the classical SBM. We then describe how it can be readily extended to the weighted case. The challenge and motivation for community identification is then introduced. We next provide a Bayesian formulation of the weighted SBM, and then review an approach using Variational Bayes (VB) to recover community identities and other parameters of interest. We then derive a Gibbs sampler and present numerical experiments comparing our sampler to the VB approach. We conclude by discussing limitations and speculate as to reasons for the disappointing performance of our sampler.

1 Introduction

The investigation of the statistical properties of networks is a quickly growing field with many important applications. While data analysis tasks in conventional statistical settings frequently require assumptions about independence, network data is fundamentally relational, as an observed network encodes relationships (edges) between units of observation (vertices).

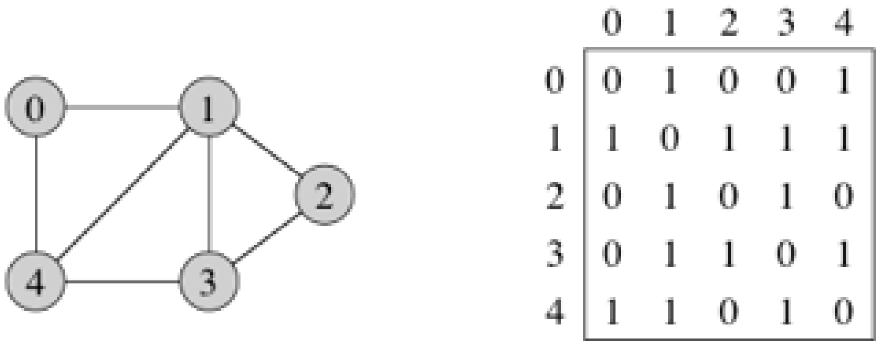
While substantial prior work has focused on the statistical properties of binary networks, many modern applications involve networks that are most naturally considered weighted. For example, in neuroscience, networks are frequently used to capture brain connectivity information, where edges represent some measure of dependence (frequently pairwise correlations of timeseries) between multiple locations in the brain, as captured with various brain imaging techniques (e.g., functional magnetic resonance imaging).

Many networks exhibit community structure, i.e., rather than treating each edge weight as following a unique distribution, one can group vertices in such a way that certain groups of edges become stochastically equivalent conditional on these groups. A natural task, then, is to infer the community labels for a given network.

This report considers the problem of community detection in the weighted network setting for a particular class of generative networks models, i.e., the stochastic block model.

2 Network Preliminaries and Notation

Let $G = (V, E)$ be a graph, where V is a set of vertices and $E \subseteq \{(v, v') : v, v' \in V\}$ is a set of edges connecting some of the vertices in V. In particular, if $(v, v') \in E$, then there is an edge linking vertex v to vertex v'. Suppose that $|V| = n < \infty$, i.e., we have n vertices. Without loss of generality, number the vertices V as $1, 2, \ldots, n$. In this case, we can represent the edge information using the adjacency matrix $A \in \{0, 1\}^{n \times n}$, where $A_{i,j} = 1 \iff (i, j) \in E$. See Figure 1 for a graphical depiction of the relationship between graphs and their representation as adjacency matrices.



	0	1	2	3	4
0	0	1	0	0	1
1	1	0	1	1	1
2	0	1	0	1	0
3	0	1	1	0	1
4	1	1	0	1	0

Figure 1: Undirected Binary Graph and Adjacency Matrix from Stack Exchange

While substantial prior work has focused on this problem in the classical setting where the network is a binary graph, i.e., edges are in $\{0, 1\}$, in this report we consider *weighted graphs*, i.e., complete graphs where each edge has an attribute in $\mathbb{R}$. We can naturally adjust the adjacency matrix such that now, $A \in \mathbb{R}^{n \times n}$, where the i, j entry now captures the *weight* of the edge linking the i'th and j'th vertices.

3 Background: Stochastic Block Models

The stochastic block model (SBM) was first introduced by Holland, Laskey, and Leinhardt [3] and is a frequently employed generative model used in the analysis of networks in a variety of settings. Under the SBM, we suppose each $v \in V$ can be assigned to one of K communities, i.e., let $\boldsymbol{z} \in \{1, 2, \ldots, K\}^n$ give community assignments. This assignment process is stochastic, with $z_i \overset{iid}{\sim} \text{Categorical}(p_1, p_2, \ldots, p_K)$. Conditional on $\boldsymbol{z}$, the edges are independent with a distribution that depends only on the community memberships of the incident nodes, i.e., $A_{i,j} \mid \boldsymbol{z} \overset{ind}{\sim} \text{Bernoulli}(P_{z(i),z(j)})$, where $P \in [0, 1]^{K \times K}$ is a matrix of parameters that govern the probability of an edge for group of dyads.

It is possible to extend the SBM to the weighted case. First, we suppose that all edges are present, i.e.,

$$E = \{(i, j) : i \neq j, \in [n]\}$$

, and as discussed above we let entries of A take values in $\mathbb{R}$. Rather than a Bernoulli likelihood for each edge's existence, we can consider a more general class of distributions, but for convenience we restrict ourselves to consideration of exponential families. Now, we can replace the matrix of parameters P with $\theta \in \mathbb{R}^{n \times n}$, and now

$$A_{i,j} | \boldsymbol{z} \overset{ind}{\sim} F(\theta_{z(i),z(j)}),$$

where F is some law parameterized by θ.

4 Bayesian Formulation

We follow a formulation laid out in Aicher, Jacobs, and Clauset [1, 2]. As discussed above, we suppose that all edge distributions come from common exponential family, and we use conjugate priors for each of these parameters, i.e.,

$$\pi_{\tau_r}(\theta_r) = \frac{1}{Z(\tau_r)} \exp(\tau_r \cdot \eta(\theta_r)).$$

We then choose a flat prior for $\boldsymbol{z}$:

$$\pi_i(z_i) = \text{Categorical}\left(\frac{1}{K}, \dots, \frac{1}{K}\right).$$

Moreover, we assume $\boldsymbol{z} \perp \boldsymbol{\theta}$. We can then write the prior as

$$\pi(z, \theta \mid \boldsymbol{\tau}) = \prod_{i}^{n} \frac{1}{K} \prod_{r}^{R} \frac{1}{Z(\tau_r)} \exp(\tau_r \cdot \eta(\theta_r))$$

. Letting $\pi^\star(\boldsymbol{z}, \theta)$ be the posterior, we have

$$\pi^\star \propto P(A \mid \boldsymbol{z}, \theta) \pi(\boldsymbol{z}, \theta).$$

With this in place, the likelihood is given by

$$P(A \mid \boldsymbol{z}, \theta) = \left[\prod_{i<j} h(A_{i,j})\right] \exp\left(\sum_{i<j} T(A_{i,j}) \eta(\theta_{z_i,z_j})\right),$$

and the posterior by

$$\pi^*(\boldsymbol{z}, \theta) = \frac{P(A \mid \boldsymbol{z}, \theta) \pi(z) \pi(\theta)}{\int_\Theta \sum_{\boldsymbol{z} \in \mathcal{Z}} P(A \mid \boldsymbol{z}, \theta) \pi(\boldsymbol{z}) \pi(\theta) d\theta}$$

5 Variational Bayes Approach

Unfortunately, the posterior described above is presents substantial computational challenges due to the combinatorics invovled in integrating out over $\boldsymbol{z}$. Aicher, Jacobs, and Clauset [1, 2] propose instead to approximate π^* by a factorizable q, i.e.,

$$q(\boldsymbol{z}, \theta \mid \boldsymbol{\mu}^*, \boldsymbol{\tau}^*) = \prod_i \mu_i^*(z_i) \times \prod_r \frac{1}{Z(\tau_r^*)} \exp(\tau_r^* \cdot \eta(\theta_r))$$

6 Sampling-Based Approach

Snijders and Nowicki [4] proposes a Gibbs sampler for the setting of the undirected SBM with binary edges. Their notation differs from (and directly conflicts with) the conventions in [1, 2], so our treatment here modifies their notation to be internally consistent. In essence, they propose a two-step sampler. First, we draw $(\mu, \theta) \sim \pi^*(X^p, y)$

7 Discussion

In this project I propose to investigate and (re)-implement procedures introduced in [1, 2] for estimation of community labels for weighted stochastic block models, possibly extending these approaches to settings with a sample of networks and nodal covariates.

These papers consider a weighted extension of the stochastic block model (SBM) [3], which is a frequently employed generative model used in the analysis of networks in a variety of settings. In particular, the SBM provides a generative model for the construction of network data. In the classic binary case with n nodes, $c \in \{1, 2, \ldots, K\}^n$ is a random vector that serves to assign each of the nodes to one of K "communities." Conditional on c, the edges of the adjacency matrix $A_{i,j}$ are independent Bernoulli; all edges in a given "block" (i.e., all those edges connecting nodes from community k to community k') are further identically distributed with probability of connection given by $B_{k,k'}$. In most settings, all that is observed is a single adjacency matrix A and the task is to infer the community memberships c and to subsequently estimate the entries of B.

The data that I consider in my research is motivated by applications in cognitive neuroscience and is related to the networks generated by the stochastic block model, but with several important extensions. First, I have been studying cases where we observe a *sample* of networks on a common node set (or where the node set can be meaningfully registered across observations. Second, the networks that I study are typically (i) dense, (ii) weighted, and (iii) signed. Finally, I consider cases where in addition to observing the adjacency matrices $A^{(i)}$, we also observe nodal attributes $X^{(i)} \in \mathbb{R}^{n \times p}$, where p is the number of covariates observed at each node.

While my current projects are largely focused on predicting observation-specific labels (y_i) using (A_i, X_i), in that work we currently consider c fixed, known, and shared across all observations. However, in future iterations of the project I intend to consider cases where c is considered unknown or possibly cases where c varies across the sample. It has been suggested to me by my advisor that a better understanding of procedures that estimate c, which frequently involve some sort of stochastic optimization, in addition to developing an implementation myself, would be a useful tool for subsequent steps in my research, and that this class project is a good opportunity to undertake this work.

Unfortunately, direct estimation of c is in most settings an NP-hard combinatorial problem, and so alternative strategies are employed. In [1, 2] the authors introduce variational Bayesian algorithms for estimating the posterior distribution of both the community labels and the parameters governing the edge distributions. A reference implementation for these approaches is provided by the authors online, and as a pedagogical project I will re-implement their technique. Depending on progress in this front, I will consider extensions that use Monte Carlo strategies from class in order to approximate for the posterior (rather than using variational methods). If this proves successful, I will explore extensions of their model (and code) to the "sample-of-networks" settings described above.

References

[1] Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset. "Adapting the Stochastic Block Model to Edge-Weighted Networks". In: *arXiv:1305.5782 [physics, stat]* (May 2013). arXiv: 1305.5782.

[2] Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset. "Learning Latent Block Structure in Weighted Networks". In: *Journal of Complex Networks* 3.2 (June 2015). arXiv: 1404.0431, pp. 221–248. ISSN: 2051-1310, 2051-1329. DOI: 10.1093/comnet/cnu026.

[3] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social Networks* 5.2 (June 1, 1983), pp. 109–137. ISSN: 0378-8733. DOI: 10.1016/0378-8733(83)90021-7.

[4] Tom A.B. Snijders and Krzysztof Nowicki. "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure". en. In: *Journal of Classification* 14.1 (Jan. 1997), pp. 75–100. ISSN: 1432-1343. DOI: 10.1007/s003579900004.