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 as an alternative to the VB approach. Unfortunately, difficulties with the implementation precluded the inclusion of numerical results, but work on this project is ongoing as community detection in the weighted network setting will support current applied research projects involving brain imaging. We conclude by discussing limitations and speculate as to challenges inherent in the approach outlined in 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.
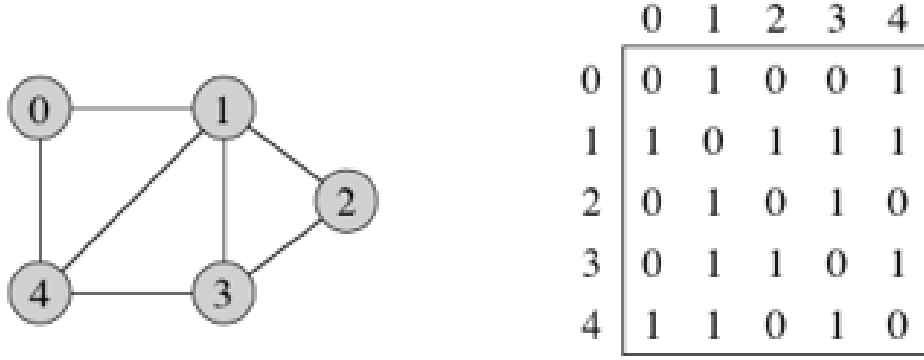
Figure 1: Undirected Binary Graph and Adjacency Matrix from Stack Exchange

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 $z_i \stackrel{iid}{\sim} \text{Categorical}(p_1, p_2, \ldots, p_K)$. Conditional on $\boldsymbol{z}$, the edges are independent with a distribution that depends only on the community memberships of the incident nodes, i.e., $A_{i,j} \mid \boldsymbol{z} \stackrel{ind}{\sim} \text{Bernoulli}(P_{z(i), z(j)})$, 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
$$A_{i,j} | \boldsymbol{z} \stackrel{ind}{\sim} F(\theta_{z(i), z(j)}),$$
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 $\boldsymbol{z}$:
$$\pi_i(z_i) = \text{Categorical}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right).$$

2

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

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

. 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 involved in integrating out over $\boldsymbol{z}$. Aicher, Jacobs, and Clauset [1, 2] propose instead to approximate π^* by a factorizable $q(\boldsymbol{z}, \theta) = q_{\boldsymbol{z}}(\boldsymbol{z})q_\theta(\theta)$, in particular

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

We wish to choose the q with minimal KL-divergence with posterior:

$$\operatorname*{argmin}_q D_{KL}(q \parallel \pi^\star) = -\int q \log \frac{\pi^\star}{q}$$

While this minimization may seem especially difficult given the unpleasant nature of the posterior π^*, there's another way. Recall that the marginal log likelihood of the data is a fixed quantity since A has been observed. Through manipulation, we obtain

$$
\begin{aligned}
C &= \log P(A) \\
&= \int_\Theta \sum_{\boldsymbol{z} \in \mathcal{Z}} q(\boldsymbol{z}, \theta)d\theta \log P(A) && \text{(Multiply by 1)} \\
&= \int_\Theta \sum_{\boldsymbol{z} \in \mathcal{Z}} q(\boldsymbol{z}, \theta) \log \frac{P(A, \boldsymbol{z}, \theta)}{P(\boldsymbol{z}, \theta \mid A)}d\theta && \text{(Conditioning tricks)} \\
&= \underbrace{\int_\Theta \sum_{\boldsymbol{z} \in \mathcal{Z}} q(\boldsymbol{z}, \theta) \log \frac{P(A, \boldsymbol{z}, \theta)}{q(\boldsymbol{z}, \theta)}d\theta}_{\mathcal{G}(q)} - \underbrace{\int_\Theta \sum_{\boldsymbol{z} \in \mathcal{Z}} q(\boldsymbol{z}, \theta) \frac{P(\boldsymbol{z}, \theta \mid A)}{q(\boldsymbol{z}, \theta)}d\theta}_{D_{\mathrm{KL}}(q \parallel \pi^*)}
\end{aligned}
$$

Finally, we can rewrite the last result as

$$D_{\mathrm{KL}}(q \parallel \pi^*) = \underbrace{\log P(A)}_{\text{Constant}} - \mathcal{G}(q),$$

thus maximizing $\mathcal{G}$ minimizes the KL divergence as desired.

We can obtain some intuition for this approach by inspecting the form of $\mathcal{G}$:

$$\mathcal{G}(q) = \int_\Theta \sum_{\mathbf{Z} \in \mathcal{Z}} q(\mathbf{z}, \theta) \log \frac{P(A, \mathbf{z}, \theta)}{q(\mathbf{z}, \theta)} d\theta$$

$$= \int_\Theta \sum_{\mathbf{Z} \in \mathcal{Z}} q(\mathbf{z}, \theta) \log \frac{P(A \mid \mathbf{z}, \theta)\pi(\mathbf{z}, \theta)}{q(\mathbf{z}, \theta)} d\theta$$

$$= \int_\Theta \sum_{\mathbf{Z} \in \mathcal{Z}} q(\mathbf{z}, \theta) \log P(A \mid \mathbf{z}, \theta) d\theta + \int_\Theta \sum_{\mathbf{Z} \in \mathcal{Z}} q(\mathbf{z}, \theta) \log \frac{\pi(\mathbf{z}, \theta)}{q(\mathbf{z}, \theta)} d\theta$$

$$= \mathbb{E}_q \left[\log P(A \mid \mathbf{z}, \theta) \right] + \underbrace{\mathbb{E}_q \left[\log \frac{\pi(\mathbf{z}, \theta)}{q(\mathbf{z}, \theta)} \right]}_{-D_{\mathrm{KL}}(q \| \pi)}.$$

In the last line, we observe that we will be choosing a q that maximizes our expected log likelihood, but that is regularized through the second term, which essentially penalizes deviation from the prior π. Aicher, Jacobs, and Clauset [1, 2] note that as with many problems in Bayesian inference, given sufficiently large data, this first term will typically dominate and overwhelm the contribution of the prior.

Aicher, Jacobs, and Clauset [1, 2] present an iterative algorithm for maximizing $\mathcal{G}(q)$. Briefly, we have that

$$G \propto \sum_r^R E_q(T_r + \tau_r - \tau_r^*)E_q(\eta(\theta_r)) + \sum_r^R \log \frac{Z(\tau_r)}{Z(\tau_r^*)} + \sum_i + \sum_{z_i \in [K]} \mu_i^* \log \frac{\mu_i(z_i)}{\mu_i^*(z_i)}$$

$$E_q T_r = \sum_{i<j} \sum_{(z_i, z_j)=r} \mu_i^*(z_i)\mu_j^*(z_j)T(A_{i,j})$$

$$E_q \eta(\theta_r) = \frac{\partial}{\partial \tau} \log Z(\tau) \mid_{\tau=\tau_r^*},$$

and by taking gradients we can obtain update rules. MATLAB code for this algorithm is provided by the authors online.

6 Sampling-Based Approach

We wished to explore a sampling-based approach in addition to the VB approach outlined above. Snijders and Nowicki [4] propose a Gibbs sampler for the setting of the undirected SBM with binary edges, which we extend to the weighted case. Their notation differs from (and directly conflicts with) the conventions in [1, 2], so our treatment here modifies their notation to be internally consistent.

First, we use a slightly extended form of the model introduced in section 4, where we will put a prior on μ. Since the likelihood of $\mathbf{z} \mid \mu$ is multinomial[1], choosing a Dirichlet prior with hyperparameter α is a natural choice that gives conjugacy. Like the model from Aicher, Jacobs, and Clauset [1, 2], our approach can be readily generalized to exponential family likelihoods for A in general, for notational compactness we consider the case where the likelihood of A is normal with known (unit) variance, i.e., $A_{i,j} \mid \theta, \mathbf{z} \sim N(\theta_{z_i,z_j}, 1)$, in which case it is natural to choose a normal distribution as the prior for θ with hyperparameters μ_0 and σ_0^2 in order to achieve conjugacy, i.e., $\theta_{i,j} \sim N(\mu_0, \sigma_0^2)$. A graphical depiction of our model is presented in Figure 2.

Adapting their approach, we arrive at two-step sampler. Assuming we index our samples by p, and supposing we have current values $\mu^{(p)}, \theta^{(p)}, \mathbf{z}^{(p)}$, we first draw $(\theta^{(p+1)}, \mu^{(p+1)}$ from the conditional distribution $(\theta, \mu \mid Z^{(p)}, A)$. Next, we iterate through the components of $\mathbf{z}$, and draw each conditional on all the other $\mathbf{z}$ (using any new draws in the present step where possible),

---

[1]or with some notational adjustment, categorical

In particular, our approach involves the following. First, $\theta \mid \mathbf{Z}, A, \mu \overset{d}{=} \theta \mid \mathbf{Z}, A$, by d-separation argument from the DAG, see Figure 2). Letting $n_{i,j}$ denote the number of edges in the i,j block and letting $t_{i,j}$ denote the sum of the edges in the i,j block (with blocks defined based on the current Z), by conjugacy we have

$$\theta_{i,j} \mid \mathbf{Z}, A; \mu_0, \sigma_0^2 \sim N\left(\left(\frac{1}{\sigma_0^2} + \frac{n_{i,j}}{\sigma^2}\right)^{-1}\right)\left(\frac{\mu_0}{\sigma_0^2} + \frac{t_{i,j}}{\sigma^2}\right)\right)$$

and this draw is straightforward.

Next, $\mu \mid A, \mathbf{Z} \overset{d}{=} \mu \mid Z$, again by DAG argument. In particular, letting $\mathbf{r}$ be a vector holding the number of vertices assigned to each of the communities according to the current Z, we have

$$\mu \mid \mathbf{Z} \sim \text{Dirichlet}(\alpha + \mathbf{r}),$$

and this draw is also straightforward.

Drawing the Z_i is slightly more involved. Because each of the Z_i can take one of k values, we evaluate the joint density for that value, and then by summing over the possible values we can get our normalizing constant. Thus, we need only a convention expression that is at least proportional to the joint density.

$$
\begin{aligned}
P(Z, \theta, \mu, A) &= P(A|Z, \theta, \mu)P(Z, \theta, \mu) \\
&= P(A|Z, \theta, \mu)P(Z|\theta, \mu)P(\theta, \mu) \\
&\propto P(A|Z, \theta)P(Z|\mu)
\end{aligned}
$$

The first time is the likelihood of our data, which is straightforward to compute, and the second term is the likelihood of Z, which is also straightforward since conditional on μ it is multinomial.

There are several problems with this approach. First, as is frequently the case with mixture models, the distribution is only identified up to some relabeling of the groups. For example, if we swap the labels in Z but make comparable changes to rows and columns of θ and to the components of μ, we will obtain the same likelihood. As a result, the sampler tends to visit combinatorically many modes (growing with K). Snijders and Nowicki [4] proposes a clever trick, but it appears to work only in the case where $K = 2$, and it is not immediately clear if it generalizes to weighted networks, wherein a symmetry argument is used in order to "un-permute" the samples and push them back onto a single mode that is identified by some ordering of the diagonal of θ. When the mode jumps around, it makes the marginal posterior of the $Z's$ meaningless, which undercuts the primary goal of this project which is community detection. Another numerical problem is that when sampling Z, the term involving the likelihood of the data is typically vanishingly small, which is a challenge for floating point precision. Our implementation instead uses the log likelihood, but this makes the derivation slightly more complicated. Finally, while I do believe that a DAG argument suffices to establish conditional independence of μ and θ (since their path is blocked by Z, which we also condition on), I was not able to thoroughly convince myself that the components of θ are conditionally independent, although my derivation and implementation implicitly makes this assumption.

R code implementing this procedure is available on github, along with a short script that conducts a proof-of-concept numerical experiment.[2]

7 Discussion

In this report, we introduced the stochastic block model, generalized it to the weighted network setting, and presented two related Bayesian formulations of this generative model with prior distributions over the

---

[2]While the code presently "works", i.e., produces results and not errors, the author believes that the current implementation is incorrect, as the results obtained are fairly nonsensical. The code will be updated as the author continues to work to implement this sampler so that it can be integrated into their broader research.
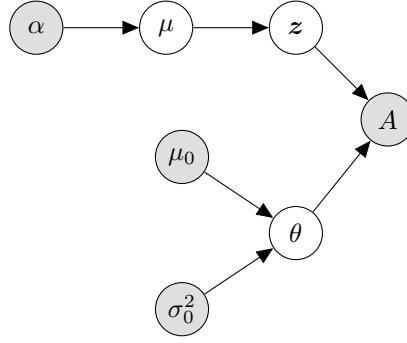
Figure 2: Graphical depiction of generative model. Grayed nodes are observed (or fixed hyperparameters), white nodes are latent.

generating parameters. We have reviewed published work implementing a Variational Bayes algorithm for community detection and detailed several of its more mystifying derivations. Subsequently, we have extended older work which presented a Gibbs sampling scheme for the classic SBM in the case where $K = 2$.

Our future work on this project, which will continue to undertake beyond this course given its relevance to our active research, will involve addressing the implementation issues discussed in section 6 above. While the lack of numerical results in this report is disappointing, it was not for lack of effort, and as discussed above a (likely incorrect) draft implementation is now posted on github along with this report.

An additional line of future work will involve extending these approaches to settings where a *sample* of networks is observed (as is typically the case in the applications I work on involving brain networks) and where it is reasonable to assume that there a common, but unknown, community structure in the population.

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.