Community Identification in Weighted Networks

Dan Kessler

April 12, 2019

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 reimplement 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.